

PCT

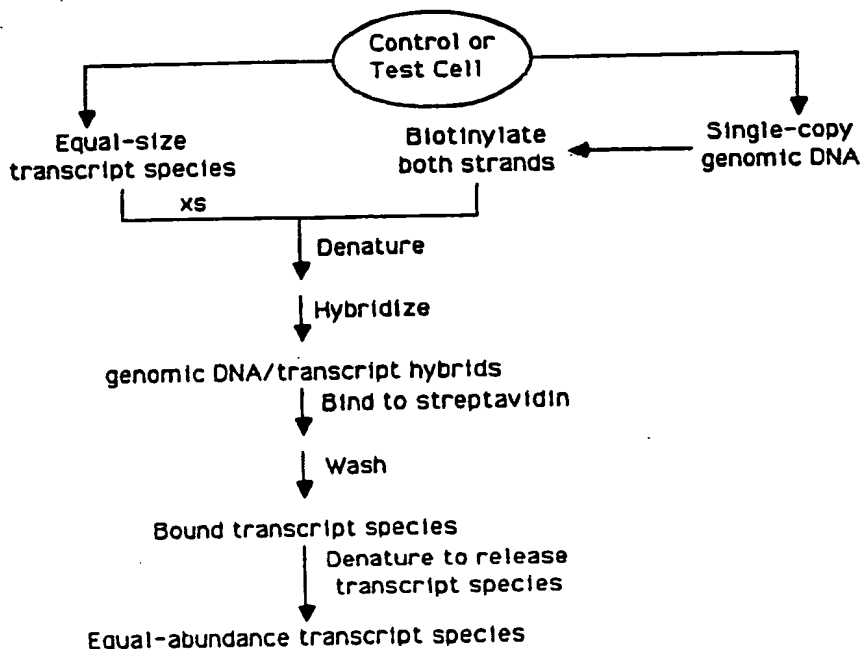
WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<p>(51) International Patent Classification<sup>4</sup> : C12Q 1/68, C12N 15/00 C07H 15/12</p>	<p>A1</p>	<p>(11) International Publication Number: WO 88/ 07585 (43) International Publication Date: 6 October 1988 (06.10.88)</p>
<p>(21) International Application Number: PCT/US88/01050 (22) International Filing Date: 1 April 1988 (01.04.88) (31) Priority Application Number: 033,980 (32) Priority Date: 2 April 1987 (02.04.87) (33) Priority Country: US (71) Applicant: GENELABS, INCORPORATED [US/US]; 505 Penobscot Drive, Redwood City, CA 94063 (US). (72) Inventors: WEISSMAN, Sherman, M. : 495 St. Ronan Street, New Haven, CT 06511 (US). REYES, Gregory, R. : 2112 St. Francis Drive, Palo Alto, CA 94303 (US). (74) Agent: DEHLINGER, Peter, J.; Ciotti &amp; Murashige, Irell &amp; Manella, 545 Middlefield Road, Suite 200, Menlo Park, CA 94025 (US).</p>		<p>(81) Designated States: AT (European patent), AU, BE (European patent), CH (European patent), DE (European patent), DK, FR (European patent), GB (European patent), IT (European patent), JP, KR, LU (European patent), NL (European patent), SE (European patent).  <b>Published</b> <i>With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i></p>

(54) Title: EQUAL-ABUNDANCE TRANSCRIPT COMPOSITION AND METHOD



(57) Abstract

A transcript composition containing approximately equal molar abundance of all of the transcript species produced by a selected genomic structure, such as total genomic DNA or isolated chromosome or chromosome fragments. Also disclosed are methods of using the composition, or compositions derived from control and test cell types, to (a) identify

***FOR THE PURPOSES OF INFORMATION ONLY***

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FR	France	ML	Mali
AU	Australia	GA	Gabon	MR	Mauritania
BB	Barbados	GB	United Kingdom	MW	Malawi
BE	Belgium	HU	Hungary	NL	Netherlands
BG	Bulgaria	IT	Italy	NO	Norway
BJ	Benin	JP	Japan	RO	Romania
BR	Brazil	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	LI	Liechtenstein	SN	Senegal
CH	Switzerland	LK	Sri Lanka	SU	Soviet Union
CM	Cameroon	LU	Luxembourg	TD	Chad
DE	Germany, Federal Republic of	MC	Monaco	TG	Togo
DK	Denmark			US	United States of America

5

EQUAL-ABUNDANCE TRANSCRIPT COMPOSITION AND METHOD1. Field of the Invention

10           The present invention relates to a composition of mRNA or DNA transcript species which are present in the composition in substantially equal molar abundance, and to methods of preparing and using the composition.

15   2. References

Anderson, M.L.M., and Young, B.D., in Nucleic Acid Hybridization: A Practical Approach, B.D. Hames and S.J. Higgins, eds., IRL Press, Oxford, (1985), 73.

20   Bodmer, W., Cold Spring Harbor Symposia on Quantitative Biology. (1986).

Britten, R.J., and Davidson, E.H., in Nucleic Acid Hybridization. A Practical Approach, B.D. Hames and S.J. Higgins, eds., IRL Press, Oxford, (1985), 3.

Britten, R.J., and Kohne, D.E., Science (1968), 161:529.

25   Davidson, E., in Gene Activity in Early Development, Academic Press, N.Y. (1976).

Engleman, E., Human Hybridomas and Monoclonal Antibodies, E. Engleman, S.K.H. Foung, J. Larrick, and A. Raubitschek, eds., Plenum Press, (1985).

30   Hames, B.D., and Higgins, S.J., eds., Nucleic Acid Hybridization: A Practical Approach, IRL Press, Oxford, (1985).

Huynh, T.V., et al, in "DNA Cloning, Volume 1", ed. D.M. Glover, Washington, D.C.: IRL Press, 1985, Chapter 2.

-2-

Kaplan, M.P., et al, J Immunol Methods (1979) 5:131.

Kauffman, R.S., and Fields, B.N., in Fundamental Virology, B.N. Fields and D.M. Knipe, eds., Raven Press, New York, (1986), 161.

5 Lewin, B., in Gene Expression 2, John Wiley and Sons, 2nd ed. (1980).

Maniatis, T., et al, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory (1982), 280.

10 Manley, J.L., et al, Proc Nat Acad Sci (USA) (1980) 77:3855.

Messing, J. in Methods in Enzymology, 10: 20-78 (1983).

Okayama, H., et al, Mol Cell Biol (1982) 2:161.

15 Stanbury, J.B., et al, eds., The Metabolic Basis of Inherited Disease, McGraw-Hill Book Company, New York, (1978).

Van Beverow, C. and I.M. Verma, Curr Top Microbio Immunol (1986), 123:73.

### 3. Background of the Invention

20 The ability to identify differences in low-abundance messenger RNAs (mRNAs) between similar or dissimilar cell types is an important area of study in human genetics. One major application is in understanding and predicting certain disease states. For example, an absent or altered mRNA coding for a  
25 specific protein in a particular cell type is often the direct cause of a hereditary disease, while the presence of an added mRNA species may signal the beginning of malignant transformation or the latent presence of an otherwise undetectable infectious agent. Although some  
30 hereditary diseases--such as sickle cell anemia, other hemoglobinopathies and the thalassemias--are due to changes in the nature or presence of high-abundance mRNAs, a large percentage of hereditary diseases have

been shown to be or are likely to be caused by the absence of or alterations in specific proteins coded for by low-abundance mRNAs. These include Lesch-Nyhan Syndrome, Hunter's Syndrome, Hurler's Syndrome,

5 Tay-Sachs Disease and adenosine deaminase deficiency, among others (Stanbury).

It is also known that several oncogenes exist whose aberrant activation leads to malignant transformation (Van Beverow), and the detection of  
10 changes in low-abundance mRNA species will have important applications in the early detection of such transformation.

Another important application of low-abundance mRNA analysis is in the diagnosis and study of low  
15 grade, slow and latent infections with viruses or other agents, especially in a tissue containing different cell types, in which only one cell type may be infected. In particular, the transcription of virus-specific mRNA(s) may be the first indication of reactivation of a latent  
20 viral infection (Kauffman).

Other important applications include, but are not limited to, the study of gene expression changes during cell activation, embryonic development or cell cycle progression, and between similar cell types of the  
25 same or closely related species.

The major problem in the detection and analysis of low-abundance mRNAs, or complementary DNAs (cDNAs) derived from such mRNA species, is interference caused by numerous high-abundance mRNAs present in the cell.  
30 In any given cell type, there may be 10,000 - 30,000 distinct mRNA species (Davidson), and these can range in concentration from several hundred thousand molecules per cell, for high-abundance species, to only a few molecules per cell, for low-abundance species.

-4-

A number of nucleic acid hybridization techniques aimed at isolating and analyzing mRNAs or their corresponding cDNAs have been developed heretofore. These techniques rely on the ability of denatured nucleic acids to reassociate in a sequence specific manner by Watson-Crick base pairing interactions (Britten). The rate at which renaturation occurs is determined by the sequence complexity of the sample, the absolute and relative concentrations of various species, the length of DNA fragments and the conditions under which the reaction takes place (Hames):

In one common hybridization method, a selected gene probe cDNA is labeled and added in single-strand form to a saturating mixture of cellular transcripts. The presence of the probe-related transcript species can be assessed by the amount of labeled probe cDNA incorporated into double-strand material. This method is generally suitable for high- and moderate-abundance transcripts, but lacks the sensitivity required for the identification and analysis of low-abundance mRNAs due to high background levels. Another limitation of this method is the requirement for large quantities of total cellular mRNA, which may be difficult to provide, particularly in clinical specimens. Moreover, the technique requires the availability of mRNA-specific probes, which means it cannot be used to study the transcript-related basis of genetic defects or other cellular conditions for which the identity of the relevant mRNA species is unknown.

Filter hybridization to a conventional cDNA library may be used to identify differences in low-abundance mRNAs (Anderson). However, the one million or more library clones needed to insure the presence of virtually all low-abundance mRNAs

effectively eliminates the utility of this technique for screening programs.

Nucleic acid subtraction techniques have also been proposed for use in studying differences in low-abundance mRNAs between related cell types (Maniatis). Here transcript preparations from two different cell types are hybridized to completion, with the unhybridized remainder being examined for the presence of species of interest. In practice, this method will also pick up transcript species whose concentrations in the two samples are different, and transcript species which are present because the hybridization reaction does not, in fact, go to completion. Therefore, the method results in a wide spectrum of transcript species with different abundance levels, and each of the species must be rechecked against the original mRNAs because of incomplete subtraction. Such subtraction techniques also lack the requisite sensitivity for the reliable detection of specific low-abundance mRNAs, and are too cumbersome and difficult for use in routine clinical applications.

#### 4. Summary of the Invention

It is therefore a general object of the invention to provide a transcript composition which substantially overcomes above-discussed prior-art limitations in isolating and analyzing low-abundance transcripts in cells.

A more specific object of the invention is to provide a composition which can be used to detect changes in the abundance and/or presence of low-abundance mRNA transcripts which occur at different cellular states and/or differentiate one cell type, or group of cells, from another cell type or group.

-6-

Still another object of the invention is to provide methods for producing such a composition.

Another general object of the invention is to provide methods for analyzing cell transcript  
5 composition, abundance, changes and/or cell-specific differences, particularly as they relate to low-abundance transcript species.

A related object is to provide such methods which are suitable for clinical application, such as in  
10 detecting markers of genetic diseases in clinical specimens.

The composition of the invention is derived from a cellular genomic structure, such as an entire cellular genome, an isolated chromosome, or a portion of  
15 chromosome, from a defined cell type or cell group. The genomic structure contains a plurality of genes  $G_i$  which are active in the defined cell type in producing messenger RNA (mRNA), at various levels of mRNA abundance. The composition includes a transcript  
20 species  $T_i$  for each  $G_i$  gene, and in substantially equal molar abundance. The transcript species may be an mRNA, or fragment thereof, or single- or double-strand cDNA, or homologous genomic DNA fragments, and the transcript species may be cloned in a suitable cloning  
25 vector.

In one embodiment, the transcript species are derived from mRNA transcripts which are within a selected size class of mRNAs, e.g., 500-2,000 base pairs. In other and preferred embodiments, the  
30 transcript species are all substantially equal sizes, typically about 300-800 base pairs, and are derived from either the 3' or 5' end regions of all of the mRNAs produced by the selected genomic structure.



-7-

The composition is prepared, according to a preferred method of the invention, by providing (a) sequences from fragments of the cellular genomic structure, in which highly repetitive genomic fragments have been substantially removed, and (b) the different-abundance cellular transcript species produced by the genomic structure. The genomic fragments are hybridized with a large molar excess of the transcript species, yielding fragment/transcript-species hybrids which can be isolated to yield the desired composition. The isolation preferably involves labeling the fragments with an affinity label, such as biotin, which permits binding to a solid support, such as an avidin-coated support, and selectively retaining hybridized transcript species by affinity chromatography.

Alternatively, an equal-abundance transcript composition can be obtained by direct hybridization of equal molar amounts of sense and anti-sense coding strands of total cellular transcript species. The opposite-strand species used in the hybridization reaction are preferably equal-size fragments derived either from the 3'- or 5'-ends of full-length transcripts. This approach is based on the more rapid (second-order) annealing rate of higher-concentration transcript species, which favors more rapid hybridization of the more abundant strands, and particularly so when all of the strands have the same length. At a selected annealing reaction time ( $C_0t$  value), the abundance of each non-annealed species will be substantially equal. When this point is reached, the non-annealed species (which are present in roughly equal abundance) are separated from annealed duplex DNA by hydroxyapatite chromatography.

The invention also includes a method of detecting mRNA transcripts which are produced by the genes of a selected genomic structure in either a test or control cell type, but not by the corresponding genes contained in the other cell type. The method uses equal-abundance transcript compositions, of the type described above, derived from both the test and control cells. Equal-abundance transcripts from one of the two cell types are labeled, e.g., by photobiotinylation, and hybridized in molar excess with the equal-abundance species from the other cell type. Unique unlabeled transcript species are isolated by affinity chromatography.

Alternatively, unique species can be identified by transferring the library equal-abundance clones from the test (or control) cell onto a filter, and hybridizing with the radiolabeled equal-abundance species from the control (or test) cell. Those clones on the filter which do not show labeled (hybridized) DNA are identified as unique to the test (or control) cell.

The invention further includes a method of detecting differences in the abundance of mRNA transcripts produced by a test cell type with respect to a control cell type. Here a control cell transcript composition of the type described above is plated on to replica filters. One of these filters is hybridized with total radiolabeled control cell transcript species, and the other, with total radiolabeled test cell transcript species. Autoradiographs of the two filters allow the density of film spots at corresponding filter positions to be compared, providing a measure of the relative transcript abundance associated with each transcript species which is common to both the control and test cells.

These and other objects of the invention will be more fully apparent when the following detailed description of the invention is read in conjunction with the accompanying drawings.

5

#### Brief Description of the Drawings

Figure 1 is a flow diagram of the steps used in preparing the equal-abundance transcript composition of the invention;

10        Figures 2A-2C illustrate three vector constructs useful in practicing the invention;

Figure 3 illustrates methods for preparing equal-abundance libraries derived from selected-size classes of mRNAs, where the transcript inserts are carried (A) in an efficient transcription vector and (B), in a vector which can be made single-stranded;

15        Figure 4 illustrates methods for preparing the equal-abundance libraries of Figure 3, where the amount of transcript material used in preparing the libraries is first amplified by an initial cloning step;

20        Figure 5 outlines steps in preparing equal-abundance cDNA libraries of 3'-end fragment transcript species according to one method of the invention, where an intermediate cloning step is not (A) and is (B) required;

25        Figure 6 outlines steps for preparing equal-abundance cDNA libraries of either 5'-end (A) or 3'-end (B) fragments according to another method of the invention;

30        Figure 7 outlines steps for preparing equal-abundance cDNA libraries of 5'-end fragments according to yet another method of the invention;

Figure 8 shows steps in producing a full-length equal-abundance cDNA library, using one of the

-10-

end-fragment equal-abundance libraries illustrated in Figures 5, 6, or 7;

Figure 9 illustrates one method for identifying transcript species which are produced by a test cell, but not a control cell;

Figure 10 illustrates a second method for identifying transcript species which are unique to a test cell;

Figure 11 outlines a method for determining the relative abundances of transcript species in control and test cells;

Figure 12 shows a method which uses a control cell equal-abundance cDNA composition prepared according to the invention, for identifying cell products which are unique to a test cell; and

Figure 13 illustrates the use of an equal-abundance cDNA composition formed from selected genomic chromosomes or chromosomal fragments, according to the invention, for determining the transcript products of the genomic structure.

### Detailed Description of the Invention

#### I. Preparing an Equal-Abundance Transcript Composition

Figure 1 illustrates steps in preparing the equal-abundance composition of the invention, according to a preferred embodiment of the invention. Briefly, genomic DNA or a selected fraction thereof (genomic structure) from a control or test cell type is isolated, fragmented, and treated to remove multiple-copy fragments. The genomic fragments are preferably labeled with an affinity label, such as biotin, to allow binding to a solid support, such as an avidin-coated support beads. These steps are indicated at the upper right in

-11-

Figure 1, and detailed below (Section IA and Examples 1 and 2).

The fragments of the genomic structure are hybridized with cellular transcript species derived from the selected cell or cell group containing the genomic structure. As used herein, the term "transcript species" refers to an mRNA transcript type or kind  $T_i$  produced by each gene  $G_i$  in the genomic structure which is active in producing mRNA transcripts in vivo. The transcript species may be a full or partial-length mRNA transcript, or a single-strand or double-strand cDNA derived from the mRNA transcript or transcript fragment, and may be derived directly from cell mRNA isolates, or from cloned cDNAs.

The transcript preparation or mixture which is hybridized with the genomic fragments can be prepared in a variety of ways. The simplest transcript preparation is total full-length mRNA isolated directly from the cell, or the corresponding cDNA. This preparation typically contains a range of transcript sizes, from a few hundred to several thousand base pairs, and a wide range of abundances, from a few copies per cell, to a hundred thousand or more copies per cell. If a large molar excess of this preparation is mixed with the labeled genomic fragments under hybridization conditions, each transcript molecule will hybridize with approximately one genomic fragment. Assuming that the average size of the genomic fragments is about the same as that of the smallest transcript size, the total number of genomic fragments available for hybridizing to each transcript species will be roughly proportional to the full-length transcript size.

To illustrate, the practical size of genomic fragments, for purposes of removing multiple-copy

-12-

genomic material, is about 300-800 base pairs. Therefore, a gene  $G_1$  whose transcript size is about 500 base pairs will be represented by about 1 genomic fragment, whereas a gene  $G_2$  whose transcript size is about 5,000 base pairs, will be represented by about 10 genomic fragments. Since the transcripts are in large molar excess of the genomic fragments (e.g., 250-fold excess), about ten times more genomic fragments are available for binding to the 5,000 base pair transcript than to the 500 base pair transcript. The resulting transcript composition--consisting of all transcript species which hybridize with the genomic fragments--would therefore contain substantially about one  $T_1$  transcript species for each ten  $T_2$  transcript species, and all other species would likewise be represented in abundances which are approximately related to transcript size. (The size dependence is actually more complicated, due to the representation of intron regions in the genomic fragments, but not the transcripts.) The resulting transcript composition may be thought of as having a size-equalized or size-normalized composition of transcript species.

Two strategies are used to reduce or eliminate the size dependence in the transcript compositions. In the first, described in Section C below, total cellular mRNAs are initially size fractionated, to produce selected size classes of transcripts, e.g., in the 500-2,000, 2,000-4,000, and over 4,000 base pair ranges. Each or a selected mRNA size class is then individually hybridized with the genomic fragments, yielding a transcript composition in which the molar amounts of the different transcript species vary within a fairly small range, e.g., one to fourfold, depending on the range of transcript sizes.

-13-

The second, and preferred approach to reducing size dependence in the transcript composition is to equalize all of the transcript sizes prior to hybridization with the genomic fragments. The equalized transcript pieces may be derived either from 3'-end regions of the transcripts, as described in Section D below, or from 5'-end pieces, as described in Section E. The equal-length transcript end pieces, when mixed with the genomic fragments under hybridization conditions, will hybridize with corresponding genomic end fragments only, and additional transcript hybrids related to total transcript length will not form. It can be appreciated that the resulting transcript composition will contain substantially equal molar amounts of each transcript species, where each transcript species is represented by a 3'- or 5'- end fragment. As defined herein, a transcript composition is also said to have substantially equal molar abundances of its transcript species if the composition of transcript species is size normalized, as defined above, and where the sizes of transcript species are within a defined size range.

Figure 1 illustrates a preferred method for carrying out the hybridization and transcript separation steps referred to above. The transcript species which are hybridized with biotinylated genomic fragments are preferably equal-size transcript species of the type indicated above, and preferably cloned 3'-end single-strand mRNA or cDNA species prepared as described in Section D below. A large molar excess of the transcript species is mixed with biotinylated genomic fragments, the mixture is denatured by heating, then cooled to allow slow annealing of the fragments with the transcript species.

After hybrid formation, the material is applied to a streptavidin column, which retains all of the biotinylated genomic fragments, including those hybridized with the transcript species. The column is washed thoroughly to remove transcript species not associated with the labeled fragments, and the transcript species are then released from the column by heat denaturation. The hybridization and transcript separation procedures just described are detailed in Example 3.

The equal-abundance transcripts are typically cloned into a suitable cloning vector to form a cDNA library. This library in turn can be used in preparing a full-length, equal-abundance cDNA library, as will be described in Section F.

#### A. Preparing Single-Copy Genomic DNA Fragments

As indicated above, the genomic DNA fragments used for hybridizing the transcript species are preferably single-copy fragments, and are derived from a selected cell or cell group whose transcript characteristics are to be examined. As defined herein, "genomic DNA structure" is intended to include total genomic DNA, or a fragment size class thereof, isolated from chromosomes, or fragments or selected regions thereof. The DNA structure includes a plurality of genes  $G_i$  which are active, in the selected cell, in producing mRNAs at various levels of mRNA abundance. The actual number of genes  $G_i$  in the structure may be relatively small (less than about 25), but typically contains hundreds or thousands of genes  $G_i$ , and the range of transcript abundances produced by the genes in the cell ranges from a few copies per cell, up to  $10^5$  or more copies per cell.



-15-

The cell from which the DNA structure is derived may be a cell line capable of sustained growth in culture, such as a variety of fibroblast or lymphocytic lines, or newly immortalized

5 B-lymphoblastoid cell lines. Sources of such cell lines, and methods for obtaining and culturing cell lines, and of immortalizing B-lymphocytes, are well known. A requirement of all such cultures, however, is that there not be any amplification or deletion of

10 genetic material ( $G_i$ ) which would lead to an altered abundance of  $T_i$ .

Alternatively, the cell source may be a cell type or group isolated from living tissue (or whole organs or entire organisms), and suspended in culture,

15 such as peripheral blood lymphocytes, or primary cultures of human embryonic lung or foreskin fibroblast cells which can be maintained and grown for a limited time in culture. Cultures of this type would have the lowest probability of developing chromosomal aneuploidy.

20 To obtain the genomic DNA structure, the cell source is fractionated, according to conventional methods, to obtain total DNA, total nuclear DNA, and/or isolated chromosomes. The isolated DNA material or structure may be further fractionated to yield

25 chromosomal regions of interest, for example defined-size restriction fragments from total genomic DNA or from one or more isolated chromosomes. Section IIC below, for example, describes a transcript composition formed from a NotI fragment of isolated

30 human chromosome 7.

DNA from the selected cell source is isolated by standard procedures, which typically include successive phenol and phenol/chloroform extractions with ethanol precipitation, according to standard procedures

(Maniatis, p. 280). Example 1A below describes the isolation of total DNA from human peripheral blood lymphocytes (PBLs) by the same procedures.

In the usual case, where the genomic material is derived from a eukaryotic cell or cell type, the DNA will consist of a relatively small percent of single-copy genes--the material of interest for purposes of the present invention-- and a major portion of repeat-sequence DNA. A preferred method for removing repeated sequences is by conventional hybridization methods which exploit the greater rate of hybridization of multiple-copy gene sequences (Britten). Briefly, in carrying out this method, DNA material is fragmented, such as by sonication or high-pressure extrusion, yielding fragments which are preferably between about 300-800 base pairs, as discussed in Example 1B. These fragments are treated under salt and temperature conditions which cause disassociation, then reannealed slowly, by dropping the temperature below the measured melting temperature  $T_m$ , the midpoint temperature of transition between single and double strands of the DNA material. Since the rate of reassociation is greater for multiple-copy fragments, the last group of fragments to reanneal will be predominantly the single-copy gene fragment material.

The annealing reaction is allowed to proceed to a predetermined  $C_0t$  (initial DNA concentration  $C_0$  times annealing time  $t$ ) value at which multiple-copy fragments are predominantly annealed and single-copy material remains single-stranded. The partially annealed reaction mixture is then separated over hydroxyapatite, which selectively binds double-strand (duplex) DNA. Additional separation between single-copy and multi-copy fragments can be achieved by repeating

-17-

the above hybridization procedure one or more times. The method is illustrated in Example 1C, and detailed in the published literature (Britten).

Alternatively, initial and/or secondary removal  
5 of multiple-copy genes may be carried out by other known techniques including hybridization and selective subtraction with known repetitive sequences such as the AluI and KpnI repeat families (Lewin). In some cases it may also be desirable to perform additional subtractions  
10 with known moderately repetitive sequences, such as histone and immunoglobulin genes and various pseudogene-families, such as nerve growth factor genes, and with known polymorphic sequences such as those coding for the antigens of the major histocompatibility complex (MHC).

15 The predominantly single-copy genomic fragments produced as above are now labeled in a manner which will permit them to be physically separated from unlabeled RNA or DNA species. The label is preferably an affinity label, such as biotin, which binds specifically and with  
20 high affinity to a surface-modified solid support, such as a solid support containing surface-bound avidin or streptavidin (Brigati, Herman). Example 2 details several methods for biotin-labeling one or both strands of duplex DNA. Streptavidin support materials are  
25 commercially available (Example 2).

As indicated above, the genomic fragments in the hybridization with transcript species serve to provide substantially equal molar quantities of each coding region of the genomic structure. Since most  
30 multiple-copy segments in a genome are non-coding, it will be appreciated that a fragment composition in which none or only a portion of the multiple-copy fragments has been removed will nonetheless provide substantially equal molar amounts of coding fragments. However, such

-18-

a multiple-copy composition would have to be added to the transcript species at much higher concentration, and therefore the specificity of the hybridization reaction would be reduced, and the reaction time would be increased over a single-copy composition. In the case where the genomic fragments are not first treated to remove multiple-copy species, the multiple-copy fragments will largely hybridize with one another and, when added to an affinity column, become bound to the column material through both strands.

#### B. Modified Cloning Vectors

This section describes three modified vectors which are useful for cloning full-length or end fragment transcript species, as will be described below. Each of the three vectors is formed by introducing one or more rare restriction sites into or adjacent the normal "cloning sites" of a conventional, commercially available cloning vector. The rare cutting site(s) allow the cloning vector with its transcript insert to be cut adjacent one or both ends of the insert, with a substantially reduced risk of cutting the transcript insert itself at an internal cutting site, which would result in the loss of a portion of the insert.

Figure 2A illustrates the modification of a pGEM-3 plasmid to form a cloning vector, designated pGEM-3/NS, which is useful for producing single-strand transcripts of cloned inserts. The pGEM-3 plasmid is commercially available from Promega Biotech (Madison, WI). As shown, this vector contains, in a 5'-to-3' direction, an SP6 RNA polymerase promoter, a polylinker cloning site region bounded by HindIII and EcoRI sites, and a T7 RNA polymerase promoter adjacent the EcoRI site. The two promoters are oriented to promote RNA

-19-

transcription of inserts cloned into the polylinker region. Thus, one strand of a DNA insert contained in the cloning site can be transcribed by an SP6 polymerase, and the other strand, by the T7 polymerase.

5 In the usual transcription procedure, the vector with the DNA insert is cut at the insert end opposite the desired promoter, so that transcription from the desired promoter terminates at the end of the transcript.

The pGEM-3 modification is aimed at replacing  
10 the cloning site region of the plasmid with a segment containing a single NotI (NI) site adjacent the HindIII (HIII) site, and a single SfiI (SI) site adjacent the EcoRI (RI) site. As illustrated in Figure 2A, this is done by cutting the plasmid with HindIII and EcoRI, to  
15 remove the polylinker region, and inserting into the cut vector an oligonucleotide containing, in a 5'-to-3' direction, a HindIII cohesive end, a NotI recognition sequence, an SfiI recognition sequence and an EcoRI cohesive end. The following oligonucleotide, which  
20 contains contiguous NotI and SfiI sites, is exemplary:

```
5'    AGCTTGCGGCCGCGGCCGGGGGGGCCG    3'
3'    ACGCCGGCGCCGGCCCCCGGCTTAA    5'
```

25

Methods for preparing oligonucleotides synthetically are referenced in the Methods and Material section below, and services which specialize in the synthesis of such molecules are available; e.g., Synthetic Genetics, Inc.  
30 (San Diego, CA) and Applied Biosystems (Foster City, CA).

After digesting the pGEM-3 plasmid with HindIII and EcoRI, the linearized plasmid is purified from the excised polylinker fragment by electroelution. The purified vector is mixed with an equimolar amount of the

-20-

oligonucleotide under conditions which favor circularization, via the oligonucleotide, and the plasmid is ligated with T4 DNA ligase conventionally. The vector construction methods follow standard procedures, such as those referenced in the Material and Methods section below. Successful recombinants are selected for ampicillin resistance on E. coli strain DH5, substantially as described in Example 4B. The ligation restores the HindIII and EcoRI plasmid sites, yielding a modified pGEM-3 plasmid, designated pGEM-3/NS. As seen in Figure 2A, the plasmid contains, in a 5'-to-3' direction, the SP6 promoter, HindIII, NotI, SfiI, and EcoRI sites, and the T7 promoter. A similar plasmid construction in which the positions of the NotI and SfiI sites are reversed can be prepared by insertion of an oligonucleotide like that above, but where the positions of the two rare sites with respect to the sticky ends are reversed.

The second vector construct, illustrated in Figure 2B, is designed for cloning applications in which cloned transcript inserts are hybridized with labeled single-strand DNA or RNA in a single-strand vector form. A preferred plasmid for modification is the Bluescribe M13+/- (M13+/-) plasmid shown in Figure 2B, and which in fact represents a pair of plasmids, designated + or -, whose origin or replication is designed to produce phage packaging of either + or - duplex insert strands, respectively. This plasmid (pair), which is commercially available from Stratagene, Inc. (San Diego, CA), contains a cloning site bounded by EcoRI and HindIII insertion sites, as shown, and an F1 origin of replication from the intergenic region of M13, which permits encapsidation of the plasmid in single-strand form in a bacterial host co-infected with

-21-

the plasmid and helper phage, as has been described (Messing). The plasmid also contains a T7 polymerase promoter adjacent its HindIII site, and a T3 polymerase promoter adjacent its EcoRI site, as shown, for  
5 transcription of a sense strand from the T7 promoter and an anti-sense strand from the T3 promoter, analogous to the pGEM-3 vector.

The modification of the M13+/- vector is carried out substantially as above, by cutting the  
10 plasmid at its unique HindIII and EcoRI sites, ligating the oligonucleotide shown above into the linearized vector, and selecting for ampicillin resistance in transformed host E. coli strain JM101. The modified plasmid, designated M13+/-/NS, includes, in a 5' to 3'  
15 direction, a T7 promoter, unique NotI and SfiI sites, and the T3 promoter.

The third modified vector, shown in Figure 2C is used in generating a HindIII/SfiI/NotI/PstI oligo dG fragment, shown at the bottom in the figure, which in  
20 turn is used in forming a 5'-end cloning vector of the type described by Okayama and Berg (Okayama), for plasmid-primed first- and second-strand cDNA synthesis.

The modified vector is derived from a pSV1932 plasmid which is available commercially from PL  
25 Biochemicals (Milwaukee, WI), and which has a HindIII/Bcl segment containing an internal PstI site and a portion of the SV40 RNA PolII promoter, as shown. The modification is aimed at introducing a single SfiI and a single NotI site immediately adjacent the Hind III site  
30 (between the HindIII and PstI sites). This is done, according to standard procedures, by cutting the pSV1932 plasmid at the unique Hind III site, and mixing the linearized vector with the following oligonucleotide:

-22-

```
5'      AGCTTGGCCGGGGGGGCGCGGCCGC      3'
3'      ACCGGCCCCCCCCGGCGCCGGCGTCGA      5'
```

which contains a cohesive HindIII 5' end followed by an  
5 SfiI site, a NotI site, and a 3' cohesive HindIII end.  
The oligonucleotide is prepared synthetically, as above,  
and is designed to regenerate only the 5' HindIII with a  
blocked 3'-end HindIII site. The vector is inserted  
into the cut HindIII site of pSV1932, ligated, and  
10 successful recombinants are selected for ampicillin  
resistance on E. coli strain DH5. The selected vector  
is vector is designated pSV/SN.

The pSV/SN plasmid is linearized by digestion  
with PstI, and treated with terminal transferase in the  
15 presence of GTP, substantially as has been reported  
(Okayama), to attach oligo dG tails to both ends of the  
vector. After removing the terminal transferase and  
GTP, the vector is cut with HindIII to release the  
desired HindIII/NotI/PstI- oligo dG fragment as shown.  
20 The use of this fragment in a procedure for cloning  
5'-end transcript fragments will be described in Section  
E below.

### C. Size-Class Equal-Abundance Libraries

25 The first equal-abundance transcript  
composition which will be considered is a size-class  
composition in which equalized transcript species are  
all derived from a selected size class of full-length  
mRNAs.

30 One method of forming this composition is  
illustrated in Figure 3, and described generally in  
Example 3. This method is generally suitable when the  
total amount of size-class mRNA available from the  
tissue or cell line of interest is sufficient to allow



the equal-abundance composition to be prepared by direct hybridization and transcript selection, using labeled genomic fragments.

It is often of interest to compare the transcript characteristics of one cell type with the same cell type in a different development or activation state, for example B-cells before and after activation with Epstein-Barr Virus (EBV). As will be seen, this application generally requires two transcript compositions--one derived from control (e.g., inactivated) cells, and the second, from test (e.g., activated) cells. Methods of maintaining a variety of cells in culture and of activating, stimulating, or otherwise altering the metabolic and biochemical behavior of cells in culture are well known for many cell systems.

Another source of mRNA is whole tissue, such as organ tissue obtained from a human or animal. The latter tissue source may be treated initially, by known techniques, to separate one cell type from other cell types present in the tissue source. Often the cell source will be from a diseased tissue, such as from a tumor tissue, or from the tissue of an individual with a known genetic disease. Methods for obtaining defined cell types or groups of cells from whole organ or tissue samples are well known.

Typically, the mRNA transcript material derived from at least about  $10^8$  cells, corresponding to a cell volume of about  $0.5 \text{ mm}^3$  of packed cell material, is sufficient. More generally, however, it may be necessary to determine experimentally for each cell or tissue type, and depending on the yield of poly A RNA achievable, whether an intermediate cloning step is required. The starting material is polyA selected mRNA

derived from the control or test cell of interest, by conventional extraction and selection procedures (Example 3A).

The mRNAs may be sized into one of a convenient number of different size classes, such as 500-2,000; 2,000-4,000; 4,000-7,000; and 7,000 and greater base pair transcript ranges. Alternatively, the transcripts can be sized into a single desired size range, e.g., 1,000-1,500 base pair transcripts, with all other size transcripts being discarded. The preferred method for sizing transcripts is by formaldehyde agarose gel electrophoresis, which allows separation into well defined size classes. Here known molecular weight RNAs are used as molecular weight markers, to gauge migration distance on the gels as a function of transcript size. The transcripts of each selected size class can be removed from the gel and purified by standard electroelution and ethanol precipitation steps, as outlined generally in Example 3B.

With continued reference to Figure 3, the sized transcripts (or corresponding cDNAs) are hybridized with labeled, single-copy genomic fragments, as outlined above, and detailed in Example 3D, to produce an equal-abundance size-class transcript composition in which each transcript species  $T_i$  is represented in molar amount substantially in proportion to its size (Example 3E). For example, assuming an average genomic fragment size of about 500 base pairs, and a transcript size class of between 500-2,000 base pairs, each of the largest transcript species would be represented by approximately four times the molar abundances of the smallest species. The size-equalized transcript species produced as above may be cloned into a suitable cloning vector by conventional methods.

-25-

One preferred cloning method, which gives directional cloning of full-copy cDNA species into an efficient transcription vector, and rare cutting sites at the transcript ends, is illustrated in Figure 3.

5 Here the transcript species are treated conventionally to produce full-copy duplex cDNAs with 5'-end hairpin loop. The duplex material is blunt-end repaired, and ligated with SfiI linkers, attaching the linkers to the duplex 3' ends. The duplex molecules are next treated  
10 with nuclease S1, to remove the 5'-end hairpins, blunt-end repaired, and ligated to NotI linkers. Digestion of the duplex molecules with NotI and SfiI yields molecules with 5'-end NotI sticky ends, and 3'-end SfiI sticky ends, as indicated. These duplex  
15 molecules are now inserted into either the pGEM-3/NS plasmid, or the M13+/-/NS vector, at the NotI and SfiI sites created in these vectors. The resulting library clones, shown at the bottom in Figure 3A and 3B, both contain the equal-abundance full-size transcript species  
20 oriented in a 5'-to-3' direction from the NotI to the SfiI sites. Preparation of the pGEM-3/NS and M13+/-/NS libraries is detailed in Examples 4A and 4B, respectively.

As indicated above, an important advantage of  
25 the pGEM-3/NS library is the ability to prepare either coding or non-coding mRNA strands efficiently. The advantage of the M13+/-/NS library is the ability to convert the plasmids to single-strand phage which are suitable for hybridization studies, as will be described  
30 below in Section II. Depending on the orientation of the F1 origin of replication, either the + or - strand of the cDNA transcript will be packaged.

The total number of recombinants which are selected in forming the transcript library is preferably

-26-

large enough to include the expected number of active genes in the gene structure of interest, and more preferably several times the number of anticipated genes. The expected number of active genes in total genomic DNA from a selected mammalian cell type is at least about 10,000-30,000 (Davidson). Since there will be a random distribution of each transcript species in the selected library, selecting at least about three clones for each expected gene guarantees that substantially all species will be represented in the library. Thus, about 100,000 transcript clones would be selected for a full-genome library which would, with high probability, include at least one transcript species for each active gene in the genome.

Since any limited size class of transcript species would contain substantially fewer transcript species than the total number of cellular transcript species, the equal-abundance libraries derived from size-class mRNAs can be made proportionately smaller than the 100,000 clone library indicated above. Similarly, in forming equal-abundance libraries derived from isolated chromosomes, or other subpopulations of the total genomic material, the total number of equal-abundance library clones required to "span" the genomic structure of interest may also be reduced considerably, as the complexity of the DNA sequence is reduced. However, for purposes of illustration, it will be assumed that total library sizes of about  $10^5$  clones are desired in all cases. Here it is noted that the 20 or so plates needed to support a library of this size is readily screened, using screening or selection procedures described in Section II below.

In cases where the total amount of cellular mRNA obtainable from control or test cells is relatively

-27-

small, e.g., where less than about  $10^8$  cells are available for RNA extraction, the preparation of an equal-abundance library further includes an intermediate cloning step to boost or amplify the total quantity of transcript material.

One intermediate cloning method is illustrated in Figure 4A. Here full-copy duplex cDNAs prepared from cellular mRNAs are equipped with 5'-end NotI and 3'-end SfiI sites, as described with reference to Figure 3, and the molecules are cloned in the pGEM-3/NS plasmid. The plasmids, after selection for successful recombinants and grown under conditions which favor plasmid production (selective amplification with chloramphenicol) are harvested, digested with SfiI, and transcribed with SP6 RNA polymerase. The resulting mRNA transcripts are coding-strand transcript species which are presumably full-length and present in approximately the same molar ratios as total cellular mRNAs, although in much greater quantity.

The clone-derived transcripts are used to prepare an equal-abundance transcript composition, which may be used to form an equal-abundance library, substantially as described above. Details of the method are given in Example 5.

Figure 4B illustrates a second intermediate cloning method for producing a size-class equal-abundance library, according to the invention. In this method, full-length duplex cDNAs derived from the total cellular transcripts are equipped with 5'-end NotI and 3'-end SfiI sticky ends, and cloned into the M13+/-/NS plasmid as above. As indicated in Section IB above, this plasmid contains an F1 origin of replication which allows encapsidation of the single-strand form of the vector when a suitable bacterial host is coinfecting

-28-

with a helper phage, such as M13 (Messing). Thus, depending on whether the + or - vector is used, a single-strand + or - form of the cloned insert can be produced readily by infecting the bacterial host that  
5 harbors the recombinant plasmids with a helper phage, harvesting the encapsidated single-strand library phage, and isolating the DNA from the phage. Details of these methods are described or referenced in Example 6.

The isolated phage DNA, which carries  
10 single-strand cDNA inserts derived from the total cellular mRNA, now amplified and presumably in about the same molar abundance as in cellular mRNAs, is hybridized with labeled genomic fragments, and separated by affinity chromatography, as above, to yield an  
15 equal-abundance composition of the phage inserts. The equal-abundance phage, after release from the affinity column, are used to produce duplex library plasmids, by direct transformation of E. coli strain JM101 and selection for ampicillin resistance, substantially as  
20 described or referenced in Example 6.

#### D. 3'-End Fragment Equal-Abundance Libraries

This section describes an equal-abundance composition and library in which the equimolar  
25 transcript species are 3'-end transcript fragments, and preferably fragments whose size range is comparable to that of the labeled genomic fragments used in preparing the composition.

Figure 5 illustrates one method for preparing  
30 the 3'-end fragment composition. The discussion will follow first the pathway at the left (Figure 5A) which is applicable when the original quantity of cellular mRNA obtained from test or control cells is sufficient for direct preparation of the composition, i.e., where

at least about  $10^8$  cells are available for mRNA isolation. Fragmenting by RNase digestion or alkaline hydrolysis are both suitable. The fragmentation treatment may be monitored, to insure optimum  
5 conditions, by agarose gel electrophoresis, as described generally in Example 7.

The 3'-end fragments from above are separated from the other mRNA fragments by oligo dT chromatography, and the isolated segments are  
10 transcribed, by standard oligo dT first-strand priming, to form single-strand cDNA. Alternatively, oligo dT priming and first-strand synthesis can be applied to the total RNA fragment mix, yielding cDNA fragments  
15 corresponding to the 3'-end, polyA containing RNA pieces only. RNA fragments can be removed by alkaline hydrolysis or RNase treatment, leaving the 3'-end cDNA fragments.

Total 3'-end, single-strand cDNAs from either of the two procedures above are now made equal abundance  
20 by hybridization with the labeled genomic fragments, in the presence of a large molar excess of the cDNA material. Since all of the 3'-end transcript fragments are about the same length, and approximately equal in length with the genomic fragment, each transcript is  
25 expected to hybridize with one or at most two genomic fragments. Assuming each genomic fragment is a single-copy species, the resulting hybridized 3'-end transcript species will then be present in molar ratios of either 1 or 2 copies per genomic gene.

30 The equal-abundance cDNA composition can then be transcribed to form double stranded duplex 3'-end fragments, and these fragments equipped with 3'-end SfiI and 5'-end NotI sticky end for cloning into pGEM-3/NS or M13+/-/NS cloning vectors, as above. Selecting

successful recombinants and forming a library of sufficient size to encompass substantially all of the expected active genes is done according to the general procedures above.

5           The pathway at the right in the figure (Figure 5B) illustrates an intermediate 3'-end fragment cloning step which is required where the amount of originally  
10           obtained cellular mRNA is low, e.g., where the source cellular material contains fewer than about  $10^8$   
15           cells. In this procedure, the 3'-end polyA mRNA or corresponding 3'-end poly dT single-strand cDNA is transcribed to form duplex cDNAs and these are equipped with 5'-end (hairpin end) NotI and 3'-end SfiI sticky ends, and cloned into either the pGEM-3/NS or M13+/-/NS  
20           cloning vectors, substantially as described with reference to Figure 3. The vectors, in turn, are used to produce equal-abundance libraries of 3'-end transcript species, substantially as described with reference to Figure 4.

25           Figure 6B illustrates a second general method for producing a 3'-end fragment equal-abundance composition and library. The method inherently involves an intermediate cloning step, and is thus especially suited to cell samples in which limited mRNA starting material is available.

30           As a first step, the full-copy mRNA starting material is transcribed to form full-length duplex cDNAs, and these molecules are end-repaired and ligated with SfiI linkers, which are added at the 3'-end of the duplex molecules only. The full-length cDNAs are now fragmented, such as by sonication, under conditions which produce duplex fragments predominantly in the 300-700 basepair size range. After repair of the fragment ends, NotI linkers are ligated onto the blunt



-31-

5' ends. Digestion with SfiI and NotI yields fragments of the following types: (a) 3'-end fragments having 5'-end NotI and 3'-end SfiI sticky ends; (b) internal fragments whose ends are both NotI sticky ends; and (c) 5' 5'-end fragments having 3'-end NotI sticky ends and 5'-end hairpins.

Selection of the 3'-end fragment types is made by cloning the fragments in a pGEM-3/NS or M13+/-/NS cloning vector cut at its NotI and SfiI sites, under conditions which favor circularization of the vector with the fragment inserts, and selecting for successful recombinants. Details of the method are described or referenced in Example 8.

The cloning vectors just described, and carrying the 3'-end fragment inserts, are manipulated to provide single-strand transcript RNA or DNA, hybridized with labeled genomic DNA, selected for equal-abundance composition by affinity chromatography, and recloned in a suitable cloning vector, substantially as described with reference to Figure 3.

One advantage of the 3'-end fragments, for use in producing an equal-abundance composition, is that fragments with co-terminal 5'-ends, but different 3' termini (i.e., mRNA transcribed in the same direction with a common 5'-end) will not hybridize with one another, or with common genomic fragments. Therefore, the potential problem of "dilution" of such common co-terminal transcripts, by binding to common genomic fragments, is eliminated.

30

#### E. 5'-End Fragment Equal-Abundance Libraries

The equal-abundance compositions described in this section are prepared by hybridizing approximately equal-size 5'-end transcript fragments with the genomic

-32-

DNA fragments. Preferred methods for preparing these 5'-end compositions employ an intermediate cloning step in which total full-length mRNAs are used to generate a library of total 5'-end fragments, similar to the second method described used in preparing cloned 3'-end transcript fragments. The one additional step which is critical to the preparation of an equal-abundance 5'-end library is aimed at isolating full-length mRNA as the substrate for the cDNA reactions. This is accomplished by one of two methods, both of which are based on affinity binding to a unique 5' terminal structure or cap added post-transcriptionally to the mature message (Lewin). The first method involves affinity chromatography with phenyl borate columns, according to published methods. The second is based on affinity binding of the mature mRNAs to anti-cap antibody.

The cloned 5'-end fragments in turn provide a source of 5'-end pieces for hybridizing with labeled genomic fragments, as above, to generate the desired equal-abundance library. Since both methods involve an intermediate cloning step, both are suitable in cases where cellular mRNA starting material is limited.

One method for generating a 5'-end fragment equal-abundance composition is illustrated in Figure 6A. As seen, the method follows many of the same steps described above with respect to Figure 6B for generating the 3'-end fragments. Specifically, the duplex cDNA fragments produced by fragmenting full-copy duplex cDNA are end repaired and ligated to SfiI linkers, equipping all of the fragment ends except the 5'-end hairpins with SfiI sites. The fragments are then digested with S1 nuclease, to remove the 5'-end hairpins, end repaired, and ligated with NotI linkers. Cleavage of the fragments with NotI and SfiI enzymes produces two

classes of fragments: (a) 5'-end fragments having 5'-end NotI and 3'-end SfiI sticky ends; and (b) fragments whose ends are both SfiI sticky ends. Selection of the 5'-end fragment types is made by cloning the fragments

5 in a pGEM-3/NS or M13+/-/NS cloning vector cut at its NotI and SfiI sites, under conditions which favor circularization of the vector with the fragment inserts, and selecting for successful recombinants. Details of the method are described or referenced in Example 9.

10 The cloning vectors just described, which carry the 5'-end fragment inserts, are manipulated to provide single-strand transcript RNA or DNA, hybridized with labeled genomic DNA, selected for equal-abundance composition by affinity chromatography, and recloned in

15 a suitable cloning vector, substantially as described with reference to Figure 4.

Figure 7 illustrates a second general method for producing a 5'-end fragment equal-abundance composition and library. In this approach, full-length

20 mRNAs are employed to generate a library of directionally oriented duplex cDNAs in the plasmid vector pSV1932, following the procedure of Okayama and Berg (Okayama). Briefly, pSV1932 is linearized by digestion with KpnI, and equipped with 5' and 3' strand

25 oligo dT tails at opposite ends of the linearized vector, by treatment with terminal transferase. The 5' strand oligo dT end is removed from the vector by digestion with HpaI, and the vector annealed through its single oligo dT tail to the poly A region of full-length

30 mRNA. The attached mRNA is now copied with reverse transcriptase, and the ends of the vector provided with 5' and 3' end oligo dC ends in the presence of terminal transferase. The resulting construct is illustrated at the top of Figure 7. The method to this point follows published methods (Okayama).

-34-

The vector construct at the top of Figure 7 is is cut with HindIII to release the 5' oligo dC end, the released segment is replaced by the HindIII/SfiI/NotI/PstI-oligo dG segment from Figure 2C, and the vector is circularized and ligated to form the plasmid shown in the middle of the figure. As seen, the resulting library vectors each contain a full-copy transcript cDNA duplex insert, oriented in a 5'-to-3' direction between HindIII, SfiI, NotI, and PstI sites at the 5' end, and a PvuII site at the 3' end. These library vectors are fragmented, e.g., by sonication, down to about 300-800 base pair size fragments, and the sonicated ends are repaired (blunt-ended) and equipped with SfiI linkers. Cleavage of the fragments with NotI and SfiI enzymes produces three classes of fragments: (a) 5'-end cDNA fragments having 5'-end NotI and 3'-end SfiI sticky ends; (b) fragments whose ends are both SfiI sticky ends; and (c) a 5'-end oligonucleotide derived from the 5'-end second strand primer. Here it is noted that the combination of both SfiI and NotI sites adjacent the 5'-end of the cDNA inserts is necessary for generating a NotI site adjacent the 5'-end of the insert and a predicted NotI/SfiI oligonucleotide vector fragment derived from the original plasmid.

Prior to cloning, a size selection by agarose electrophoresis and gel electroelution is performed to isolate appropriately sized 5'-end cDNAs and exclude the NotI/SfiI oligonucleotide fragments. Selection of the 5'-end fragment types is made by cloning the fragments in a pGEM-3/NS or M13+/-/NS cloning vector cut at its NotI and SfiI sites, under conditions which favor circularization of the vector with the fragment inserts, and selecting for successful recombinants. Details of the method are described or referenced in Example 10.

#### F. Preparing Full-Length Equal-Abundance Library

In some applications, it is convenient to have full-length copies of the equal-abundance transcript species, rather than 3'-end or 5'-end transcript fragment species prepared as above. One application of the equal-abundance compositions, for example, is in identifying mRNA species which are unique to a particular cell type or state (Example IIA). Here it is generally useful to be able to isolate the full-length cDNA species either directly or by hybridization in a single-step with a unique 3'-end or 5'-end fragment species. Another application which requires full-copy transcripts involves in vitro protein synthesis of transcript species for purposes of analyzing transcript products (Section IIC). One further use is the expression directly in eukaryotic cells of the full-length species. Such an application may be applied to the full-length species cloned into the pSV/SN vector (Figure 7).

In preparing the full-copy composition and library, the cloned equal-abundance 5'-end or 3'-end fragments are used as probes for total, full-length cellular mRNA transcripts, or their corresponding single-strand cDNAs. The preferred method, however, involves combined polyA selection and use of the 5'-end equal-abundance library as probe to insure isolation of a full-length transcript. Full-length transcripts which hybridize with the probes are separated from non-hybridized material. This is done, for example, by labeling the equal-abundance transcript end-fragments with an affinity label, such as biotin, using the probes to select full-length transcripts, and separating the probe-hybridized transcripts from non-hybridized material by affinity chromatography on a matrix

-36-

containing avidin. The bound full-length transcripts are then released and cloned to form a full-length equal-abundance library by the above methods.

Figure 8 illustrates the method generally, as it is applied to 3'-end or 5'-end fragment equal-abundance libraries contained in pGEM-3/NS or M13+/-/NS libraries. The particular vectors which are shown at the top in the figure are 5'-end fragment libraries such as generated according to the methods of Figure 6A or 7. The strategy when using a pGEM-3/NS cloning vector is to generate non-coding fragment transcript or transcript species, i.e., species which are capable of hybridizing with cellular mRNAs. This can be done, as above, by cutting the pGEM-3/NS cloning vectors at the 5'end NotI site, and transcribing the non-coding cDNA strand with T7 polymerase. The non-coding transcripts can be photobiotinylated by published methods. Alternatively, the cloning vector can be used to generate coding RNA transcripts, and these can then be used to select equal-abundance first strand cDNA for library construction.

To use an end-fragment M13+/-/NS cloning library, the library plasmids are made single-stranded, by infection of the plasmid-containing host with an M13 helper phage. The single-strand encapsidated phage are then isolated and treated to release the single-strand DNA, as above. The phage DNA is photobiotinylated, as above.

The biotinylated end-fragment single-strand phage DNA is hybridized with a large molar excess of full-length cellular mRNAs (using - strand phage), or the corresponding single-strand cDNAs (using + strand phage), and an equal-abundance full-length transcript composition is prepared as above. The composition may

be cloned, according to procedures described generally in Figure 3. Additional details of the method are provided in Example 11.

5    G.   Direct Hybridization Methods

          The methods for preparing equal-abundance libraries described in Sections IA-IF are based on hybridization of total transcript species with genomic fragments, and preferably single-copy genomic  
10   fragments. This section describes a second general method for preparing an equal-abundance composition. The method involves direct hybridization of complementary strands of cDNA derived from the total cellular transcripts of the cell of interest. The  
15   hybridization is carried out under concentration and temperature conditions which allow the bulk of high- and moderate-abundance, but not the low-abundance species, to hybridize with one another. That is, the hybridization reaction is carried out to a  $C_0t$  value  
20   at which transcript species which are present at a relatively low concentration (corresponding approximately to the concentration of the lowest-abundance species) are in a predominantly (50%) non-annealed form.

25           The hybridized duplex species are then separated from non-annealed molecules, e.g., by hydroxyapatite chromatography. The non-annealed molecules are present in substantially equal molar amounts, corresponding approximately to the  
30   concentration of the lowest-abundance species. Alternatively, one of the strand mixtures can be labeled, such as by biotinylation, and the non-annealed molecules of the other strand separated by affinity chromatography, e.g., with avidin column support material.

-38-

The complementary strands which are used in the hybridization are preferably equal-size 3'-end or 5'-end fragments generated as above. The advantage of equal-size fragments is that the rate of annealing of any transcript species, and therefore the relative concentration of that species with respect to all other species, is substantially independent of transcript size. The complementary, equal-size single-strand material which is annealed may be prepared by excising cloned duplex library fragments from one of the total 5'-end or 3'-end transcript libraries described in Sections ID and IE above. In this embodiment, the excised library inserts are denatured by heating, then slowly annealed to a  $C_0t$  value at which the different species are substantially equalized in molar concentrations.

In a second embodiment, the complementary sense and anti-sense strands are individually generated from the pGEM-3/NS library vector of total equal-size transcripts, as described above. The two populations of transcript strands, are then annealed as above, until the transcript species are substantially equalized in concentrations.

In yet another embodiment, biotinylated sense or anti-sense transcript from pGEM-3/NS may be combined with M13/-/NS or M13+/NS libraries, respectively, for direct hybridization and the resulting equal-abundance single-strand phage recovered by direct transformation of appropriate bacterial hosts.

The selected  $C_0t$  value which maximizes the extent to which all species are represented in substantially equal molar amounts, with a minimum loss of any transcript species, can be calculated from the estimated concentration in the mixture of the



-39-

lowest-abundance species. Specifically, the  $C_{ot}$  value selected is that which "precedes" the hybridization of any of the lowest-abundance species, as determined from the initial concentration of such a lowest-abundance species. After the initial hybridization and separation steps, the isolated non-annealed material is preferably carried through one or more additional hybridization/separation cycles, to further equalize the concentration of all of the species. The final equalized composition can be cloned in a suitable vector, such as pGEM-3/NS, as above.

Since the direct hybridization method just described does not require single-copy genomic DNA fragments, it is somewhat easier to prepare than the composition formed by hybridization to genomic fragments. However, the method has two limitations not shared by the genomic-fragment approach. First, the equal-size end-fragment species are likely to contain sequences (either 5'-end common untranslated leader or 3'-end polyA sequences) which are common to defined subsets or many if not all of the species. It would therefore, in practice, increase the difficulty in selecting a  $C_{ot}$  value to maximize equal-abundance representation.

Secondly, it is inherently more difficult to achieve true equal-abundance concentrations of the many transcript species in the direct hybridization method. This limitation reflects the differences in the range of concentrations of desired coding species which are present in genomic fragments versus total cellular transcripts. In genomic fragments, the desired single-copy species are all present at about one copy per cell, whereas repetitive genomic fragments are generally present at relatively high copy numbers. Thus

-40-

a fairly sharp separation between repetitive and single-copy coding sequences can be made in the hybridization step used to remove repetitive DNA. Accordingly, the equal-abundance composition, which is formed by hybridization to the genomic fragments, has essentially the same uniformity of concentration among the different transcript species. By contrast, the total transcript material used in forming the equalized composition in the second method contains a continuum of concentrations between lowest- and highest-abundance species. It is accordingly more difficult to achieve (or identify) a  $C_0t$  value at which the concentration of non-annealed molecules is about the same for all species.

15

## II. Utility

### A. Identifying Unique mRNA Species

Information about the presence or absence of low-abundance mRNAs is of interest in understanding the etiology of disease processes as well as fundamental cellular events relating to induction, infection, differentiation, and the like. In particular, information about the induction or repression of unique species of mRNAs would aid in (a) understanding the basis of various disease states at the gene level, (b) developing new methods for detecting cancerous or precancerous conditions, (c) diagnosing, studying, and isolating latent virus infections, and (d) studying changes in gene expression which occur during cell induction, activation, cell cycle progression, or the like.

30

In many of these cases, it is desired to identify unique mRNA species which are present in a

-41-

test, but not a control cell, or inversely, are present in the control, but not the test cell. As the term is used herein, control cell is intended to mean a reference cell against which changes in the transcript composition of the test cell are measured. For example, in studying changes in cell type which occur as a result of viral infection, embryogenic change, activation, or the like, the control cell and test cell are typically a common cell type or common cell group, before and after the cell event of interest.

One method of identifying unique mRNA species, according to the invention, is the subtraction method illustrated in Figure 9. The particular method illustrated is designed for identifying one or more unique transcript species which are present in test cell, but not control cell transcripts. It will be appreciated that transcript species which are unique to the control cell can be identified by a similar method in which the roles of the control cell and test cell equal-abundance libraries are reversed.

The method requires initially the preparation of an equal-abundance library for both the test and control cells. Where both libraries are formed from a common genomic structure, such as the total genomic material from a common cell type, the libraries can each be produced from a common labeled genomic fragment preparation, in combination with the cellular mRNA preparation from either control or test cells. The only requirement is that both equal-abundance libraries be prepared by substantially the same method, and in particular, that the cloned equal-abundance fragments from one library be capable of hybridizing with the corresponding fragments in the other library.

-42-

The libraries shown in Figure 9 are 5'-end fragment pGEM-3/NS libraries produced according to methods described above. One of the libraries, and preferably the control library, is manipulated to produce non-coding transcript species, by cutting the library plasmids with NotI and transcribing the fragment insert with T7 polymerase. The other library is similarly manipulated to produce the coding strand of the corresponding test cell fragment inserts, by cutting the plasmids with SfiI, and generating coding-strand transcripts in the presence of SP6 polymerase.

To identify unique test cell transcripts the control cell non-coding strand RNA fragments (or the corresponding cDNAs) are biotinylated, as above, and annealed in large molar excess with the test cell RNA fragments (or the corresponding cDNAs). Those test cell transcripts which hybridize with the control cell transcripts, i.e., those transcripts that are common to both cells, are removed by affinity chromatography, yielding only those test cell species which are not present in the control cell. These may now be used as a hybridization probe (after end labeling) to identify the unique transcript species present in that cell. Probes generated by these procedures will have greater sensitivity than those isolated by standard subtraction procedures. Example 13 below illustrates the method, as it is applied to identifying and isolating transcripts which are unique to EBV-activated peripheral blood lymphocytes (PBLs).

Figure 10 illustrates a second method for identifying transcripts which are unique to a test cell. The method involves first blot-transferring the plated library vectors from a test cell equal-abundance library onto a nitrocellulose filter, culturing the blot

transfer colonies to amplify plasmids within individual clones, fixing plasmid DNA to the filter, then hybridizing with radiolabeled mRNA or cDNA obtained from the control cell equal-abundance library vectors.

5 Labeled mRNA transcripts, produced, for example, by transcribing from a pGEM-3/NS cloning vector in the presence of radiolabeled ribonucleoside triphosphates, are preferred, since non-hybridized RNA can be removed by hydrolysis.

10 Development of the filters against X-ray film shows radiolabeling at all test cell clones corresponding to control cell transcripts. By comparing the pattern of radiolabel spots with the positions of the test cell clones, those test cell transcripts which  
15 do not hybridize with control cell species can be identified. The identified test cell clones, such as those indicated at A and B in Figure 10, are preferably recloned on fresh medium, reblotted on filter paper, and confirmed for non-hybridization with labeled control  
20 cell transcripts. The method is illustrated, for detecting unique transcript species related to EBV activation in peripheral blood lymphocytes, in Example 14.

Unique transcript(s) identified and isolated by  
25 the methods just described may be cloned, used in an in vitro protein synthesizing system to analyze unique protein products of the test cell, and/or radiolabeled and used for probing and isolating "unique" test cell genes, and their corresponding full-length cDNA clones  
30 from appropriate libraries.

It can be appreciated that the equal-abundance composition permits identification of low- to very low-abundance unique transcript species, due to a reduction of background levels by several orders of magnitude.

-44-

compared with conventional subtraction methods. The lower background is due both to the lower relative concentration of control cell fragments which are required to hybridize with the test cell transcript species, and to the much greater relative concentration of low-abundance species present in the test cell composition.

In the filter-hybridization method, the lower background is due to the much lower levels of radiolabeled species needed to detect species common to both control and test cells. In addition, this method would be highly impractical without an equal-abundance library, due to the very large number of library transcripts which would have to be examined.

15

#### B. Measuring Transcript Abundance

In many cell systems, changes in transcript composition between control and test cells are expected to involve changes in the levels of existing transcripts, rather than the appearance of new species or the loss of existing transcript species. The method of the present section is designed for detecting such transcript-level changes. The method is based on differential levels of binding of total mRNAs from control and test cells to the DNAs of an equal-abundance library.

The method, as outlined in Figure 11, involves plating of the equal-abundance transcript library onto two filters, one which will function for control cell hybridization, and the other, for test cell hybridization, after the plasmid DNA material is fixed to the filters.

Total cellular mRNA is isolated as above, i.e., using oligo dT chromatography to isolate total polyA

-45-

RNAs. These can be labeled by polynucleotide kinases after limited base hydrolysis. Alternatively, total cellular mRNAs isolated from the two cell types can be reverse transcribed in the presence of radiolabeled nucleotide triphosphates, to produce labeled cDNA. Each of the labeled mRNA (or corresponding cDNA) preparations is added to a filter under hybridization conditions, to identify the corresponding library clones, in proportion to the total number of copies of each species originally present in the mRNA preparation. The method assumes that total number of copies of each equal-abundance library vector is equal to or greater than the total number of mRNA molecules in the highest-abundance cellular mRNA species. For this reason, the filters are preferably prepared by growth on agar plates, as above, and under growth conditions which favor large copy number of the library vectors in plated bacteria.

After hybridization, the filters are washed, and if necessary, treated with ribonuclease to remove non-specific background associated with the RNA probe. The filters are developed against an X-ray film. The two plates in Figure 11 illustrate typical autoradiograms which are observed. Here each circle represents an individual recombinant colony from the plated library, and the density of dot shading within each circle represents the relative numbers of labeled mRNA molecules which have hybridized with each colony. In actuality, each plate would typically contain up to 5,000 or more colonies. The method is illustrated generally in Example 15.

As seen, the number of cellular transcripts which binds to all but two of the library vectors is about the same in control and test filters. In one of the library vectors, indicated by arrow A, the number of

-46-

transcripts is substantially increased in the test cell, and in the other, indicated by arrow B, it is substantially reduced.

5 The colonies corresponding to those which show a change in mRNA abundance, between control and test cell, can be picked from the original library plate, replated, and retested by hybridization with labeled cellular mRNAs, to confirm the correlation between specific library transcripts and changes in the  
10 abundance of cellular mRNA. The isolated recombinant itself might also be used as a hybridization probe against mRNAs from test and control cells to confirm its reduced or increased level of expression.

15 The feasibility of the present method is due to the relatively manageable number of recombinant clones making up the equal-abundance library which are examined for hybridization with cellular mRNAs. Where the control cell library represents the entire genome, the total number of distinct library recombinants needed to  
20 guarantee representation of nearly all cellular transcripts is about 100,000, as discussed above. Assuming a cell density of about 5,000 per plate, the entire screening procedure can be limited to only 20 sets of filters.

25

### C. Identifying Transcript Products

Another general application of the invention is in identifying total or unique transcript products associated with a selected genomic structure, such as  
30 total genomic DNA, isolated chromosomes, or large genomic fragments.

In the simplest case, an equal-abundance library is used to generate full-length equal-abundance mRNAs from a selected cell. This may be done, as



-47-

indicated above, by employing a 5'-end fragment or 3'-end fragment equal-abundance library to isolate by hybridization, an equal-abundance composition of full-length cellular mRNAs. Alternatively, full-length coding transcripts can be generated from a full-length equal-abundance library, prepared as in Section IF. Since the full-length transcripts are present in substantially equal molar amounts, in vitro protein synthesis in a conventional protein synthesizing system produces substantially equal numbers of all transcript protein products. These, in turn, can be displayed by two-dimensional electrophoresis, yielding a pattern of protein bands representing virtually all the protein products of the cell. The pattern gives information not available from conventional transcript preparations in that (a) proteins normally present in high-abundance do not mask nearby protein bands and (b) proteins normally present in small abundance are present in detectable amounts. The technique therefore is limited only by the number of proteins which co-migrate, due to similar size and charge.

The method can be extended to identifying cellular products which are unique to a test cell. This is done by comparing the pattern of bands in the 2-D gel electrophoresis patterns of control cell and test cell equal-abundance transcripts, as illustrated in Figure 12.

Figure 13 illustrates methods for generating equal-abundance transcript products from defined genomic structures. The first structure is a fraction which is substantially enriched for human chromosome 7, according to well known methods. A DNA library containing such material may be purchased commercially from the American Type Culture Collection (Rockville, MD). The fragments prepared with the enriched-for chromosome, when

hybridized with total cellular transcripts, yield an equal-abundance of all of the transcripts which are actively expressed by chromosome 7. The equal-abundance library, in turn, is used to produce an equal-abundance composition of full-length transcripts which, in an in vitro protein synthesizing system, yields approximately equal molar amounts of all of the protein products produced by genes actively expressed in human chromosome 7.

10           The size of the genomic structure can be narrowed still further, for example, to study the gene products of a selected size fragment of an isolated chromosome, such as chromosome 7. Figure 13 shows a method for preparing a library of this type. Briefly, 15 an equal-abundance library for chromosome 7 is prepared as above, and an equal-abundance library for an isolated sized genomic fragment, such as a NotI segment ( $N_i-N_j$ ) of chromosome 7 is similarly produced. These libraries may be used to select their respective 20 equal-abundance transcripts, which after in vitro translation, can be compared by 2-D gel electrophoresis. mRNA transcripts produced by in vitro transcription of non-coding and coding strands from these two separate respective libraries may also be 25 hybridized, as above, and non-hybridized material, corresponding to non-overlapping transcripts, removed by affinity chromatography, if the selecting transcript segments have previously been labeled by biotin. The  $C7/N_iN_j$  equal-abundance transcripts can be used for 30 identifying full-length cDNA clones for in vitro transcription and translation of the NotI segment specified proteins.

The following examples illustrate methods for producing size-class, 3'-end fragment, 5'-end fragment,

and full-length equal-abundance libraries, according to the invention, and to methods for using the libraries to analyze differences in transcript quantities and types between non-activated and EBV-activated peripheral blood lymphocytes. The examples are intended to illustrate preferred methods of preparation and particular uses of the composition of the invention, but are in no way intended to limit the scope of the methods or applications to other cell types or other lymphocyte states.

#### Materials and Methods

pGEM-3 is obtained from Promega Biotech (Madison, WI); Bluescribe M13+/- and helper M13 phage, from Stratagene (San Diego, CA); and E.coli strain DH5, and E. coli strain JM101, from Bethesda Research Labs (Bethesda, MD).

Terminal transferase (calf thymus), alkaline phosphatase (calf intestine), polynucleotide kinase, Klenow reagent, and S1 nuclease are all obtained from Boehringer Mannheim Biochemicals (Indianapolis, IN); SP6 and T7 polymerase, from Promega Biotech; and proteinase K, RNase and DNase, from Sigma (St. Louis, MO).

NotI, SfiI, HindIII, EcoRI, KpnI, PstI, HpaI, T4 DNA ligase and T4 DNA polymerase are obtained from New England Biolabs (Beverly, MA); oligo dT primer and oligo dA and oligo dT cellulose, from PL Biochemicals (Milwaukee, WI); Chelex-100, from Bio-Rad (Richmond, CA); Sephadex G-50, from Pharmacia (Piscataway, NJ); streptavidin agarose, from Bethesda Research Labs (Bethesda, MD); and photobiotin from Clontech Labs (Palo Alto, CA).

-50-

Synthetic oligonucleotides for vector modifications to introduce NotI and SfiI linkers are prepared by either the phosphotriester method as described by Edge, et al, Nature (supra) and Duckworth, et al, Nucleic Acids Res (1981) 9:1691 or the phosphoramidite method as described by Beaucage, S.L., and Caruthers, M.H., Tet Letts (1981) 22:1859 and Matteucci, M.D., and Caruthers, M.H., J Am Chem Soc (1981) 103:3185 and can be prepared using commercially available automated oligonucleotide synthesizers. Alternatively, custom designed synthetic oligonucleotides may be purchased, for example, from Synthetic Genetics (San Diego, CA). Kinasing of single strands prior to annealing or for labeling is achieved using an excess, e.g., approximately 10 units of polynucleotide kinase to 1 nmole substrate in the presence of 50 mM Tris, pH 7.6, 10 mM MgCl<sub>2</sub>, 5 mM dithiothreitol, 1-2 mM ATP, 1.7 pmoles  $\gamma$ 32P-ATP (2.9 mCi/mmole), 0.1 mM spermidine, 0.1 mM EDTA.

Site specific DNA cleavage is performed by treating with the suitable restriction enzyme (or enzymes) under conditions which are generally understood in the art, and the particulars of which are specified by the manufacturer of these commercially available restriction enzymes. See, e.g., New England Biolabs, Product Catalog. In general, about 1  $\mu$ g of plasmid or DNA sequence is cleaved by one unit of enzyme in about 20  $\mu$ l of buffer solution; in the examples herein, typically, an excess of restriction enzyme is used to insure complete digestion of the DNA substrate. Incubation times of about one hour to two hours at about 37°C are workable, although variations can be easily tolerated. After each incubation, protein is removed by extraction with phenol/chloroform, and may be followed

-51-

by ether extraction, and the nucleic acid recovered from aqueous fractions by precipitation with ethanol (70%). If desired, size separation of the cleaved fragments may be performed by polyacrylamide gel or agarose gel electrophoresis using standard techniques. A general description of size separations is found in Methods in Enzymology (1980) 65:499-560.

Restriction cleaved fragments may be blunt ended by treating with the large fragment of E. coli DNA polymerase I (Klenow) in the presence of the four deoxynucleotide triphosphates (dNTPs) using incubation times of about 15 to 25 min at 20 to 25°C in 50 mM Tris pH 7.6, 50 mM NaCl, 6 mM MgCl<sub>2</sub>, 6 mM DTT and 0.1-1.0 mM dNTPs. The Klenow fragment fills in at 5' single-stranded overhangs but chews back protruding 3' single strands, even though the four dNTPs are present. If desired, selective repair can be performed by supplying only one of the, or selected, dNTPs within the limitations dictated by the nature of the overhang. After treatment with Klenow, the mixture is extracted with phenol/chloroform and ethanol precipitated. Treatment under appropriate conditions with S1 nuclease results in hydrolysis of any single-stranded portions of DNA. In particular, the nicking of 5' hairpins formed on synthesis of cDNA is achieved.

Ligations are performed in 15-50 µl volumes under the following standard conditions and temperatures: for example, 20 mM Tris-Cl pH 7.5, 10 mM MgCl<sub>2</sub>, 10 mM DTT, 33 µg/ml BSA, 10 mM-50 mM NaCl, and either 40 µM ATP, 0.01-0.02 (Weiss) units T4 DNA ligase at 14°C (for "sticky end" ligation) or 1 mM ATP, 0.3-0.6 (Weiss) units T4 DNA ligase at 14°C (for "blunt end" ligation). Intermolecular "sticky end" ligations are usually performed at 33-100 µg/ml total DNA

concentrations (5-100 nM total end concentration). Intermolecular blunt end ligations are performed at 1  $\mu$ M total ends concentration.

In vector construction employing "vector fragments", the vector fragment is commonly treated with bacterial alkaline phosphatase (BAP) or calf intestinal alkaline phosphatase (CIP) in order to remove the 5' phosphate and prevent self-ligation of the vector. Digestions are conducted at pH 8 in approximately 10 mM Tris-HCl, 1 mM EDTA using about 1 unit per  $\mu$ g of BAP at 60°C for one hour or 1 unit of CIP per  $\mu$ g of vector at 37° for about one hour. In order to recover the nucleic acid fragments, the preparation is extracted with phenol/chloroform and ethanol precipitated. Alternatively, religation can be prevented in vectors which have been double digested by additional restriction enzyme digestion and separation of the unwanted fragments.

20

### Example 1

#### Preparation of Single-Copy Genomic DNA

##### A. DNA Isolation

Peripheral blood lymphocytes (PBLs) are derived from normal individuals and T cells are removed by Ficoll-Hypaque gradient (Kaplan, M.E., & Clark, C., 1979, J. Immunol. Meth. 5, 131-135). The chromosomal DNA is isolated by proteinase K digestion in the presence of 1.5 % sodium dodecyl sulfate (SDS), and 50 mM EDTA, pH 7.5, followed by successive phenol and phenol/chloroform (1:1) extractions, according to standard procedures (Maniatis). The DNA is redissolved in 0.15 M potassium phosphate buffer, pH 7.0 (PB) and passed over a Chelex 100 column to remove metal ions.

-53-

DNA concentration is determined by absorbance at 260 nm. The purity of the material is confirmed by spectrophotometric studies on melting (Britten). The material is precipitated with ethanol, and stored as a  
5 70% ethanol precipitate at -70° C until used.

#### B. DNA Fragmentation

An appropriate amount of DNA from above is collected by centrifugation and dissolved in 10 ml of  
10 0.06 M Na acetate to a DNA concentration of about 2 OD<sub>260</sub> units/ml. The DNA solution is then diluted to 30 ml with glycerol, giving a final solution which is 0.02 M Na acetate and about 66% glycerol. This material is placed in a 50-ml high-speed blender, and cooled by  
15 immersing the sides of the blender in a dry-ice/ethanol bath. Blending is begun as the solution cools, before it becomes too viscous. The material is blended at 50,000 rpm for 30 minutes. Two volumes of cold ethanol are added to the blended solution, and the material is  
20 allowed to stand in the freezer for two hours. The DNA precipitate is collected by centrifugation at 10,000 g for 15 minutes.

The range of DNA fragment size, which is preferably about 200-800 bases, is confirmed by agarose  
25 gel electrophoresis, according to standard procedures.

#### C. Removing Repetitive-Sequence DNA

The DNA-fragment sample is dissolved in 0.12 PB, 0.2 mM EDTA. Repetitive-sequence DNA is removed by  
30 standard hybridization methods which are detailed in the literature (Britten). Briefly, the DNA is raised to about 10°C above the melting temperature ( $T_m$ ), as determined for example by absorption at OD<sub>260</sub>. In the buffer used above, the  $T_m$  is between about 80-90°C.

The material is then cooled slowly to about 25°C below the  $T_m$ , and allowed to anneal to a  $C_0t$  value (mole/liter x sec) of about 100, at which the repeat-sequence material is predominantly in reannealed form, and the non-repetitive fraction, in denatured form. This duplex material is separated from single-strand DNA by hydroxyapatite (HAP) chromatography, according to standard procedures (Britten). Briefly, HAP is suspended in 0.15 PB, 2 mM EDTA, and poured into a water-jacketed column maintained at the reannealing temperature. After washing the column with several volumes of the reannealing buffer, the DNA material is loaded onto the column and the single-strand material eluted with several volumes of the buffer. This material is combined, and precipitated with cold ethanol, as above.

The precipitated single-strand material is redissolved in annealing buffer, and the entire separation procedure repeated, except that the reannealing is performed at a temperature about 10°C below the above  $T_m$  value.

#### Example 2

##### Preparation of Biotinylated Genomic DNA

Double-stranded, single-copy genomic DNA from Example I is biotinylated according to one of five methods detailed below. The biotinylated nucleotides used are Bio-11-dUTP (Brigati) which has an 11-atom linker arm separating the biotin and the pyrimidine base, and Bio-19-SS-dUTP (Herman) which has a 19-atom linker containing a disulfide bond.  $^{32}P$ -labeled dNTPs are included when monitoring of the various steps of the method is desired.



#### A. Nick-Translation

A typical reaction, carried out in 60  $\mu$ l final volume, contains 1  $\mu$ g DNA in 50 mM Tris-Cl pH 7.5, 10mM MgSO<sub>4</sub>, 0.1 mM DTT, 100 mM of each of the following  
5 nucleotides: dATP, dGTP, and Bio-11-dUTP or Bio-19-SS-dUTP, 5  $\mu$ Ci of [ $\alpha$ -<sup>32</sup>P] dCTP (Amersham, specific activity 3,000 Ci/mmol), 30 U DNA polymerase I, and 27 pg/ml DNase I. The reaction mixture is incubated at 14°C for one hour, stopped by addition of  
10 EDTA to 10 mM and heated at 68°C for 5 min. Labeled DNA is recovered by chromatography over Sephadex G50 equilibrated and eluted with 10 mM Tris-Cl, pH 7.5/1 mM EDTA (T.E.). When large amounts of DNA are required, two to three nick-translations are run in parallel and  
15 loaded onto one column to obtain a concentrated DNA solution.

#### B. Tailing by Terminal Transferase

This procedure is used only after the DNA is  
20 first treated to produce 3' protruding ends (Maniatis). The reaction mixture consists of 1  $\mu$ g DNA in 100 mM potassium cacodylate (pH 7.2), 2 mM CoCl<sub>2</sub>, 0.2 mM DTT, 100  $\mu$ M Bio-11-dUTP, 50  $\mu$ Ci [ $\alpha$ -<sup>32</sup>P] dCTP, and 20 U terminal transferase, added last. After incubation at  
25 37°C for 45 min, an additional 20 U of enzyme is added and the incubation repeated. The reaction is terminated by EDTA added to 10 mM, the DNA is recovered as described above, precipitated with ethanol, washed with 70% ethanol and resuspended in 50  $\mu$ l of T.E.

30

#### C. Labeling by T4 DNA Polymerase Replacement Reaction

The reaction contains 1  $\mu$ g of DNA in 33 mM Tris-OAc (pH 7.9), 66 mM NaOAc, 10 mM MgOAc, 0.5 mM DTT, 0.1 mg/ml BSA, and 0.5 U T4 DNA polymerase. After

-56-

incubation at 37°C for 7 minutes. dATP, dGTP, and Bio-11-dUTP are added to a final concentration of 150  $\mu$ M. dCTP is added to 10  $\mu$ M. 50  $\mu$ Ci of [ $\alpha$ -<sup>32</sup>P] dCTP (3000 Ci/mmol), and TrisOAc, NaOAc, MgOAc, BSA, and DTT are added to maintain previous concentrations. This reaction is incubated at 37°C for 30 min, then dCTP is added to a concentration of 150  $\mu$ M, and the reaction incubated for an extra 60 min at 37°C. The reaction is stopped by addition of EDTA to 10  $\mu$ M. heated at 65°C for 10 min, chromatographed and processed as described before.

#### D. Klenow Fill-In Reaction

This is carried out following standard protocols (Maniatis); incubation is at room temperature for 15 min.

#### E. Labeling by Photobiotinylation

This is carried out by standard procedures, as outlined in the protocol supplied by the manufacturer (Clontech, Palo Alto, CA).

### EXAMPLE 3

#### Preparation of Equal-Abundance Size-Class Composition: Method 1

##### A. Isolation of B-Lymphocyte mRNA

RNA is isolated from B-cells prepared as in Example 1, according to standard procedures (Maniatis, p.187), which use vanadyl ribonucleoside complexes to inhibit RNase. The total RNA preparation is fractionated by oligo dT chromatography, also according to well-known procedures (Maniatis, p. 211) yielding a polyA mRNA fraction.

#### B. Size Fractionation of mRNA

The polyA mRNA preparation from part A is fractionated by electrophoresis on 1% agarose, using glyoxal and dimethylsulfoxide to denature RNA (Maniatis, p. 150). A standard plot of log RNA size as a function of migration distance is prepared using standard RNA size standards. Three size fractions of RNA are collected: 500-2,000; 2,000-4,000; and greater than 4,000 base pairs. The RNA is eluted from the three gel regions, by performing phenol extractions on the frozen gel slices, and then collected by ethanol precipitation.

#### C. Preparation of Single-Strand cDNA

The polyA mRNA from part B, from the 2,000-4,000 size class, is used to obtain single-strand cDNA transcripts according to the method of Maniatis, et al (supra). Briefly, a portion of the polyA RNA is treated under appropriate buffer conditions with reverse transcriptase in the presence of poly dT primer, and the four nucleotide triphosphates. The complex is treated with base to destroy the remaining mRNA, and the single-strand cDNA is isolated by ethanol precipitation.

#### D. cDNA Hybridization with Single-Copy Genomic DNA

The single-strand cDNA from part C is dissolved in 0.15 M PB, 2 mM EDTA, and mixed with a solution of the single-copy biotinylated genomic DNA from Example 2, at a relative concentration of about 250 OD<sub>260</sub> units cDNA to 1 OD<sub>260</sub> unit biotinylated genomic DNA, where the OD measurement for the genomic fraction is determined in the denatured state.

The combined fractions are heated to about 10°C above the T<sub>m</sub> (Example 1), until the genomic DNA has been completely denatured, as determined by the

hyperchromic effect at  $OD_{260}$ . The material is then cooled to about 25°C below the  $T_m$  and the reannealing reaction is allowed to proceed until to a  $C_0t$  value of about 5,000, or until the reannealing process, as monitored by hyperchromic effects at  $OD_{260}$ , stabilizes.

E. Separation of Size-Class Equal-Abundance Transcripts

A 1 ml silanized syringe plugged with silanized glass wool is packed with 0.3 ml streptavidin-agarose and washed with 0.15 PB, 2 mM EDTA. The hybridization mixture from 3D is loaded onto the column which is then washed with several volumes of the hybridization buffer, to remove non-hybridized cDNA.

The column is then heated to about 10°C above the  $T_m$  value (Example 1) for about 10 minutes, then washed with heated buffer (at the same  $T_m + 10^\circ\text{C}$  temperature) to elute the desired equal-abundance single-strand cDNA. The cDNA material which is eluted is cooled to 4°C and precipitated overnight with ethanol at -20°C.

Example 4

Preparation of Equal-Abundance Size Class Libraries

A. Equal-Abundance pGEM-3/NS Library

The precipitated equal-abundance first-strand cDNA from Example 3E is taken up in 10 mM Tris-HCL, pH 8.3 containing .15 M KCl and 10 mM  $MgCl_2$ , and converted to duplex cDNA as above. The full-copy cDNAs are blunt ended and ligated at their free (3') ends with SfiI linkers. The duplex cDNA is cut with nuclease  $S_1$  to cleave the molecules at their 5' ends, repaired with Klenow reagent and ligated with NotI linkers. The duplex molecules are digested with NotI, and SfiI, to

-59-

remove redundant end linkers, yielding duplex molecules with 5' NotI and 3' SfiI sticky ends. The mix is heat treated to denature the restriction enzymes.

pGEM3 plasmid is modified to contain internal NotI and SfiI cloning sites, substantially as described in Section IB. The modified plasmid, designated pGEM-3/NS is digested with NotI and SfiI to open the vector at its unique NotI and adjacent SfiI sites. The linearized plasmid fragment is isolated by electroelution after agarose gel electrophoresis and then treated with alkaline phosphatase prior to mixing with the NotI/SfiI cDNA fragments from above, under conditions which promote circularization of single plasmid fragments, and ligated to form circularized plasmids. The plasmid includes, in a 5'-to-3' direction, an SP6 promoter, a unique NotI site, the full-copy cDNA insert, a unique SfiI site, and the T7 RNA polymerase promoter. The circularized plasmid is selected on E. coli strain DH5, and successful recombinants are selected for ampicillin resistance. The cell density of the plating step is such as to yield about 5,000 colonies per plate, on a total of about 20 plates.

#### 25 B. Equal-Abundance M13+/-/NS Library

The precipitated equal-abundance first-strand cDNA is treated as in Example 4A to produce double-strand equal-abundance cDNAs with 5'-end NotI and 3'-end SfiI sites.

30 M13+/- is modified to contain internal NotI and SfiI cloning sites, substantially as described in Section IB. The modified plasmid, designated M13+/-/NS, is digested with NotI and SfiI to open the vector at its unique NotI and adjacent SfiI sites. The linearized

-60-

plasmid fragment is isolated by electroelution after agarose gel electrophoresis and then treated with alkaline phosphatase prior to mixing with the NotI/SfiI cDNA fragments from above, under conditions which  
5 promote circularization of single plasmid fragments, and ligated to form circularized plasmids. The plasmid includes, in a 5'-to-3' direction, a T7 polymerase promoter, a unique NotI site, the full-copy cDNA insert, a unique SfiI site, and a T3 polymerase promoter. The  
10 circularized plasmid is selected on E. coli strain JM101, and successful recombinants are selected for ampicillin resistance. The cell density of the plating step is such as to yield about 5,000 colonies per plate, on a total of about 20 plates.

15

#### Example 5

##### Equal-Abundance Size-Class Composition: Method 2

This method is suitable for cell systems in which limited amounts of cellular mRNA are available, requiring an initial transcript cloning step to generate  
20 amplified amounts of total transcript species.

Total full-length mRNA from Example 3A is reverse-transcribed to form full-length, duplex cDNAs according to standard procedures, using oligo dT priming  
25 for first-strand synthesis. The full-copy cDNAs are equipped with 5'-end NotI and 3'-end SfiI sites, as in Example 4B, and these fragments are inserted into the NotI/SfiI site of pGEM-3/NS, also as in Example 4B.

Successful recombinant plasmids, selected for  
30 ampicillin resistance on E. coli strain DH5, are treated with SfiI, to open the plasmids at the 3' end of the inserts. The linearized vector is mixed with SP6 RNA polymerase in the presence of ribonucleosides, under conditions specified by the manufacturer (Promega

-61-

Biotech, Bulletin # 001), giving mRNA transcription from the SP6 promoter of the 5' strand, and yielding coding RNA transcripts.

5 The now amplified coding RNA is transcribed to single-strand cDNA, as above, using oligo dT priming, and hybridized with labeled genomic fragments from Example I. A 250 fold molar excess of the single-strand cDNA is hybridized with the biotinylated genomic DNA fragments from Examples 1 and 2, and separated from  
10 non-hybridized cDNA by affinity chromatography on a streptavidin column, as in Examples 3D and 3E. The hybridized cDNA released from the column, which is the desired equal-abundance composition, is made double stranded, equipped with 5'-end NotI and 3'-end SfiI  
15 sticky ends, and cloned into a pGEM-3/NS or M13+/-/NS cloning vector, as in Example 4B or 4C, respectively.

#### Example 6

##### Equal-Abundance Size-Class Composition: Method 3

20 This example describes an alternative method for preparing an equal-abundance transcript composition when limited amounts of cellular mRNA are available.

Total full-length mRNA from Example 3A is transcribed to form full-length, duplex cDNAs as above.  
25 The full-length cDNAs are equipped with 5'-end NotI and 3'-end SfiI sites, as in Example 4B, and these fragments are inserted into the NotI/SfiI site of M13+/-/NS, as in Example 4C.

30 E. coli strain JM101 harboring the recombinant M13+/-/NS plasmids is infected with the M13 helper phage (VCS-M13 helper phage supplied by the manufacturer), the latter permitting encapsidation of the M13+/-/NS single-strand DNA derived from the plasmids in the infected cells. Coinfection and culture conditions are

performed according to published methods. The encapsidated recombinant phage are now isolated from the culture, and single-strand DNA material prepared by conventional procedures (Messing).

5           The single-strand DNA is hybridized with labeled genomic fragments from Example I. A 250 fold molar excess of the single-strand cDNA is hybridized with the biotinylated genomic DNA fragments from Examples 1 and 2, and then separated from non-hybridized  
10 cDNA by affinity chromatography on a streptavidin column, as in Examples 3D and 3E.

          The hybridized phage material which is released from the column contains the desired equal-abundance transcript species. The single-strand phage is now  
15 converted back to its plasmid form by transforming E. coli strain JM101. Transformed colonies are then selected for ampicillin resistance provided by the M13+/-/NS plasmid. The density of cells is such as to give about 5,000 colonies, on a total of 20 plates.

20

#### Example 7

#### Preparation of an Equal-Abundance

#### 3'-End cDNA Library: Method 1

#### 25   A. Preparation of 3'-End cDNAs

          PBL mRNA is isolated as in Example 3A, and suspended in 20 mM Tris-HCl, 1 mM EDTA, pH 7.0 at 4°C. To the RNA solution is added RNase A (1 unit/ml), and the mixture is incubated at 10°C under conditions which  
30 produce RNA fragments predominantly in the 300-500 base pair regions. The reaction conditions may be established by digestion at 10°C for increasing time periods, and monitoring the size distribution of the RNA with agarose gel electrophoresis.



The fragmented RNA is extracted with phenol/chloroform, precipitated with ethanol, and redissolved in 20 mM Tris-HCl, 0.5 mM NaCl, 1 mM EDTA, 0.1% SDS, pH 7.6. The material is then fractionated by oligo dT column chromatography, by standard procedures (Maniatis, p. 197) to isolate 3' end fragments containing polyA. The 3' fragments are primed with poly dT and copied, as in Example 3, to produce 3'-end single-strand cDNA.

**B. Preparation of 3'-End Equal-Abundance Library**

A 250 fold molar excess of the 3'-end cDNAs from above are mixed with the biotinylated genomic fragments of from Examples 1 and 2, and the two DNA fractions are hybridized and bound on a streptavidin column, as described in Examples 3D and 3E. After washing the column to remove non-bound (abundant) cDNAs, the equal-abundance cDNA species are eluted by heating. The eluted material is made double stranded (Maniatis, p.214), and equipped with 5'-end NotI and 3'-end SfiI linkers, and inserted into pGEM-3/NS as in Example 4B. Successful recombinants are selected on E. coli strain DH5, at a cell density of about 5,000/plate, on a total of 20 plates.

**Example 8**

**Preparation of a 3'-End Equal-  
Abundance Library: Method 2**

Total mRNA from Example 3A is taken up in 10 mM Tris-HCL, pH 8.3 containing .15 M KCl and 10 mM MgCl<sub>2</sub>, and converted to duplex full-length cDNAs (containing a 5'-end hairpin), using oligo dT first-strand priming, as above. The 3'ends of the full-copy cDNAs are repaired

with Klenow reagent, and ligated with SfiI linkers (Example 4B). The cDNAs are fragmented by sonication, until fragment sizes predominantly between about 300-700 base pairs in length are obtained. The size distribution of the fragments as a function of sonication time can be followed by gel electrophoresis. The staggered ends of the fragments are repaired with Klenow reagents, and the fragments are ligated with NotI linkers, as above. Digestion of the fragments with both NotI and SfiI yields 3'-end fragments with 3'-end SfiI and 5'-end NotI sticky ends. All of the other fragments contain either NotI sites at both ends or NotI/hairpin opposite ends (the 5'-end fragments).

The 3'-end fragments are inserted into the NotI and SfiI sites of pGEM-3/NS as in Example 4B, and successful recombinants are selected for ampicillin resistance. A total of about  $10^5$  clones are selected, as above.

20

Example 9Preparation of a 5'-EndEqual-Abundance Library: Method 1

Total cellular mRNA prepared as above is further selected for full-length intact mRNA molecules. This is accomplished by performing oligo dT selection, as performed above, to select for post-transcriptional 3'-end processing, followed by further isolating the intact mRNA species by a procedure which is specific for intact processed (capped) 5' ends (Lewin). The latter method uses chromatography on phenol-boronic agarose (Manley), or affinity chromatography based on RNA binding to anti-cap (processed 5' end) antibody.

-65-

The double-strand cDNA fragments produced from such doubly selected mRNAs are then sonicated prior to blunt-end repair by Klenow reagent and ligation with SfiI linkers, as above. The fragments, which include

5 5'-end fragments having hairpin ends, are treated with nuclease S1 and ligated with NotI linkers as in Example 4B. Digestion of the fragments with both NotI and SfiI yields 5'-end fragments with 5'-end NotI and 3'-end SfiI sticky ends. All of the other fragments contain SfiI

10 sites at both ends.

The 5'-end fragments are inserted into the NotI and SfiI sites of pGEM-3/NS as in Example 4B, and successful recombinants are selected for ampicillin resistance. Successful recombinants are grown in liquid

15 culture in the presence of chloramphenicol, to enhance plasmid replication, and the plasmids are isolated from the cells according to standard procedures.

The isolated pGEM-3/NS plasmids are digested with SfiI, and treated with SP6 polymerase, as in

20 Example 5. The resulting RNA fragments are tailed with poly A (Maniatis). Oligo dT priming is now used to form single-strand cDNAs and these are hybridized with the labeled genomic fragments from Examples 1 and 2, and processed to form an equal-abundance cDNA composition as

25 in Examples 3D and 3E. The equal-abundance cDNAs are made double-stranded, equipped with 5'-end NotI and 3'-end SfiI sticky ends, and cloned into pGEM-3/NS as in Example 4B to form the desired 5'-end fragment library.

30

#### Example 10

#### Preparation of 5'-End Total and Equal-Abundance Fragment Libraries: Method 2

##### A. Preparing a Full-Length cDNA Library

Total polyA RNA prepared as in Example 9 is used to obtain a full-length cDNA library according to

-66-

the method of Okayama and Berg. This method differs from the usual cDNA cloning method in that (1) the plasmid vector DNA functions as the primer for the synthesis of the first cDNA strand, and (2) the second strand is prepared by priming with an oligo dG tailed fragment. Briefly, the pBR322 plasmid is opened at its unique KpnI site, with addition of oligo dT linkers to the cleaved ends. The vector is now cut at its HpaI site to remove the 5'-end oligo dT linker, and the large plasmid fragment isolated. The plasmid fragment with a 3' oligo dT linker is annealed to the polyA RNA, and the first strand cDNA is produced by copying the RNA in the presence of reverse transcriptase and the four deoxyribonucleotides, followed by addition of oligo dC tails to the single-strand cDNA using terminal transferase. The plasmid is now digested with HindIII to remove a HindIII/HpaI segment which has terminal dC base pairs.

The HindIII treated plasmid is separated from the HindIII/Hpa fragment and ligated with the HindIII/NotI/SfiI/PstI-oligo dG fragment produced substantially in accordance with Section IB, under conditions which favor circularization of the plasmid with the insert. Ligation yields a plasmid with a full-copy transcript insert bounded at its 5'-end by a HindIII, NotI, SfiI, and PstI sites, and at its 3'-end by a PvuII site, as illustrated at the center in Figure 7. The plasmid, after ligation, is added to a liquid culture of E. coli DH5, the culture is grown for 1 hour, then switched to ampicillin. Addition of chloramphenicol is used to amplify plasmid growth in the bacteria.

Full-length cDNAs cloned into the pSV/SN vector may be transfected directly into eukaryotic cells (e.g.,

COS7) to promote expression of the insert utilizing the eukaryotic transcriptional elements provided by the SV40 sequences in the vector.

5    B.   Preparing a Total 5'-End Fragment Library

Plasmid from the transformed bacteria from above are isolated by well-known methods (Maniatis) and the plasmid is fragmented by sonication into sizes of between about 300-500 base pairs, as in Example 8. The  
10   staggered ends of the plasmid fragments are blunt-end repaired, and ligated with SfiI linkers as in Example 4B, then cut with SfiI to remove redundant linkers and release plasmid-derived segments from the 5'-ends of the 5'-end insert fragments. The fragments are then treated  
15   with NotI to create a NotI sticky end at the 5'-ends of the 5' NotI/SfiI cDNA fragments, which are isolated from the SfiI/NotI linker segment by agarose gel electrophoresis prior to cloning into pGEM-3/NS as in Example 4B, with selection for ampicillin resistance.  
20   The p-GEM-3/NS library vectors are used as in Example 9 to generate an equal-abundance 5'-end transcript library.

Example 11

Preparing a Full-Length Equal-Abundance Library

25

The 5'-end fragment equal-abundance library from Example 9 or 10 is digested with NotI and SfiI to release the 5-end fragment inserts, and the total DNA is biotinylated by nick translation, as in Example 2A.

30

Total full-length mRNA from Example 9 is used to prepare full-length single-strand cDNA, as in Example 3C. The full-length, single-strand DNA material is added in 250-fold molar excess to the biotinylated 5'-end equal-abundance fragments, and the mixture is

-68-

hybridized, as in Example 3D. Single-strand material which hybridizes to the 5'-end fragments is separated by affinity chromatography. The isolated single-strand cDNA transcript material is made double stranded, and equipped with 5'-end NotI and 3'-end SfiI sticky ends as in Example 4B, and cloned into the NotI/SfiI site of a pGEM-1/NS vector, as in Example 4B, or into the NotI/SfiI site of a M13+/-/NS vector, as in Example 4C.

10

Example 12Preparing an Equal-Abundance Library  
from EBV-Activated PBLs

PBLs from a normal individual are isolated as in Example 1. The isolated B cells are cultured at  $10^7$ /ml for 1 hr at 37°C in IMDM medium containing 10% concentrated EBV supernatant from the marmoset line B95-8 (Engleman, p 454). The infected cells are washed twice and then transferred to a 75 cm<sup>2</sup> tissue culture flask, at a density of about  $5 \times 10^4$  cells/ml in IMDM medium with 10% fetal calf serum. Transformed colonies are evident at 1 week and the culture is maintained for 21 days at 37°C under 95% O<sub>2</sub>/5% CO<sub>2</sub>.

After culturing, the cells are washed two times with wash buffer (150 mM NaCl, 50 mM Tris, pH 8.3, 5 mM EDTA and 50 mM freshly added  $\beta$ -mercaptoethanol), with low-speed centrifugation to pellet the cells. The pelleted cells are extracted conventionally for mRNA (Maniatis). The total polyA RNA fraction from the EBV-activated cells is prepared as in Example 3A.

The total polyA RNA fraction from EBV-activated PBLs is used to make full-length, double-strand cDNAs, and these are used, in conjunction with the biotinylated genomic fragments from Examples 1 and 2, to produce a

total 5'-end fragment library in pGEM-3/NS, following the general procedures in Example 9.

Example 13

Identification of RNA Sequences Unique  
to EBV-Activated B-Lymphocytes: Method 1

5                   pGEM-3/NS 5'-end fragment equal representation  
libraries from B-lymphocytes not subjected to EBV  
transformation (normal-cell library) are prepared as  
10 described in Examples 9 or 10, and 12, respectively.  
The normal-cell library vectors are linearized at the  
NotI site and incubated with T7 RNA polymerase, in the  
presence of all four ribonucleotide triphosphates, to  
generate the non-coding RNA strand of the library insert  
15 (Figure 8). The RNA is photobiotinylated according to  
published techniques (Example IIE).

The activated-cell library is linearized with  
SfiI, and similarly reacted with SP6 RNA polymerase and  
ribonucleotide triphosphates, to generate the coding  
20 RNAs transcribed from the 5'-end fragment library  
inserts. This RNA preparation is mixed with about a 10  
fold molar excess of the biotinylated non-coding RNA  
from above, and the two RNA fractions are hybridized  
under slow annealing conditions. The reaction is  
25 carried to a  $C_0t$  value of about 5,000 and/or may be  
monitored at  $OD_{260}$  to determine virtual reaction  
completion. The reaction mixture-- which contains  
non-hybridized biotinylated RNA, biotinylated  
RNA/activated-cell RNA hybrids, and unhybridized unique  
30 activated-cell RNA-- is fractionated by affinity  
chromatography, using a streptavidin column. The  
initial eluate (non-bound material) contains the desired  
coding mRNA which is unique to EBV-activated cells.

-70-

The RNA fragments from above (the transcript fragments which are unique to the activated cells) are tailed with poly A at the 3' ends and these are reverse transcribed to form duplex cDNAs by first-strand oligo dT priming. The duplex molecules are equipped with 5-end NotI and 3'-end SfiI sticky ends, and cloned into the the modified transcription vector pGEM-3/NS as in Example 4B.

10

Example 14Identification of RNA Sequences Unique to EBV-Activated B-Lymphocytes: Method 2

The test cell 5'-end fragment equal-abundance library from Example 12 is plated, as described, on about 20 plates, at about 5,000 cells/plate. The colonies are replica plated onto nitrocellulose filters, and plasmid DNA fixed to the filters according to known procedures (Maniatis, p.316)

Control cell 5'-end fragment equal-abundance library plasmids from Example 9 are treated with SfiI and transcribed with SP6 in the presence of all four ribonucleotide triphosphates, including <sup>32</sup>P-labeled UTP, to form radiolabeled RNA fragments.

The radiolabeled control cell probe fragments are added to the nitrocellulose filter, under hybridization conditions (Maniatis, p326). The filters are then washed, and placed in contact with X-ray film, for radiolabel development. The filter spots which do not show radiolabeling represent colonies whose library plasmids are unique to EBV-activated cells. These colonies may be picked from the original plates for eventual confirmation of their unique expression in the test cell (EBV-activated B-lymphocytes) by dot blot hybridization against RNA isolated from activated and



-71-

non-activated B-lymphocytes. Alternatively, the unique activated RNA transcripts from Example 13 may be subjected to limited alkaline hydrolysis and end-labeled with polynucleotide kinase for use as a hybridization probe against the test cell 5'-end fragment equal-abundance library.

#### Example 15

##### Determining the Relative Abundance of Test and Control Cell Transcripts

Each plate of the control cell equal representation library (total of 20 plates) from Example 9 is replica plated on each of two detergent free nitrocellulose filters, and the plasmid DNA is fixed to the filters as in Example 14. A total of 40 filters (2 per plate) are prepared.

Total mRNA is isolated from non-activated B-lymphocytes (control cells) and from EBV-activated B-cells (test cells) as in Examples 3A and 12, respectively, and each RNA fraction is selected for full-length message by oligo dT and phenyl borate chromatography, as in Example 9.

The respective control and test cell RNAs are subjected to limited alkaline hydrolysis and then end-labeled with polynucleotide kinase for use as hybridization probes against a replica filter set. All of the filters are developed against X-ray film for a period which is found to give good differential labeling among the spots on the filters.

The extent of hybridization associated with each filter spot is estimated qualitatively, or can be quantified, for example, by a digital optical reader which is designed to output the coordinates of each spot, and the density of dark spots (radioactive decay) associated with each spot.

The density of spots on each control cell filter is compared with that from the corresponding test cell filter, to identify clones which have hybridized a substantially different amount of probe, and hence display a differential transcript abundance and level of expression. These clones are picked from the corresponding control cell plate and plated for rescreening to confirm their differential hybridization. The same experiment performed on an equal-representation library from test cells will identify expressed transcripts either unique to the activated (EBV-transformed) B-cell or exhibiting differential abundance.

While preferred methods, uses and embodiments have been described herein, it will be apparent to those in the field that various changes and modifications can be made, and the invention applied to a variety of cell systems, without departing from the scope of the invention.

25

30

IT IS CLAIMED:

1. A transcript composition derived from a  
5 cellular genomic structure containing a plurality of  
genes  $G_i$  which are each active in a defined cell type  
in producing messenger RNA transcripts, at various  
levels of messenger RNA abundance, said composition  
comprising a transcript species  $T_i$  for each  $G_i$  gene,  
10 and in substantially equal molar abundance.
2. The composition of claim 1, wherein the  
transcript species are derived from the entire genome of  
a given cell type.  
15
3. The composition of claim 1, wherein the  
transcript species are derived from a selected  
chromosome or chromosome fragment of a given cell type  
in a defined state.  
20
4. The composition of claim 1, wherein the  
transcript species are selected from the group  
consisting of messenger RNA transcripts, and single- and  
double-strand cDNA.  
25
5. The composition of claim 4, wherein the  
transcript species are cDNAs which are cloned as inserts  
in a suitable cloning vector.
- 30 6. The composition of claim 1, wherein the  
transcript species are derived from genes  $G_i$  whose  
messenger RNA transcripts are all within a defined size  
range.

7. The composition of claim 1, wherein the transcript species are all substantially equal size nucleic acid fragments derived from the 3'-end fragments of said messenger RNA transcripts.

5

8. The composition of claim 1, wherein the transcript species are substantially equal size nucleic acid fragments derived from the 5'-end fragments of the messenger RNA transcripts.

10

9. The composition of claim 1, wherein the transcript species are full-length cDNAs derived from full-length messenger RNAs.

15

10. A method of preparing a composition of transcripts species which are (a) each derived from one of a plurality of genes  $G_i$ , in a cellular genomic structure, all of which genes are active in a defined cell type in producing cellular messenger RNA transcripts, at various levels of messenger RNA abundance, and (b) present in substantially equal molar abundance, said method comprising

20

providing a collection of fragments of the cellular genomic structure,

25

providing cellular transcript species derived from the different-abundance messenger RNA transcripts,

mixing the genomic fragments with a molar excess of the cellular transcript species under conditions which promote hybridization between the fragments and homologous transcript species.

30

isolating the fragment/transcript species hybridization products formed by said mixing, and recovering the transcript species from the hybridization products.

-75-

11. The method of claim 10, wherein the collection of genomic fragment provided has been substantially depleted of repeat-sequence genomic fragments, the genomic fragments are labeled with an affinity label which permits binding to a solid support, and said isolating includes binding the fragment/transcript- species hybrids to the solid support.

12. The method of claim 10, for use in preparing a composition in which all of the transcript species are within a selected size range, wherein providing the cellular transcript species includes obtaining total cellular transcripts, and fractionating the total transcripts into the selected size range, prior to said mixing.

13. The method of claim 10, for use in preparing a composition in which all of the transcript species have substantially the same size, wherein providing the cellular transcript species includes obtaining total cellular transcripts, fragmenting the transcripts into a predetermined size range and isolating those transcript fragments containing only polyA regions, prior to said mixing.

14. The method of claim 10, for use in preparing a composition in which all of the transcript species have substantially the same size, wherein providing the cellular transcripts includes obtaining total cellular transcripts, fragmenting the transcripts into a predetermined size range and isolating those fragments containing only 5'-end regions.

15. The method of of claim 10, for use in preparing a composition in which all of the transcript species have substantially the same size, wherein providing the cellular transcript species includes  
5 obtaining total cellular messenger RNAs, attaching each RNA to a poly-dT sticky end of a linearized vector, using the attached transcript to produce a duplex DNA copy of the transcript attached to the vector at its 3' end, circularizing the vector to attach the 5' end of  
10 the duplex DNA copy adjacent a selected marker and a known restriction site in the vector, fragmenting the vector into fragments smaller than about 1,000 base pairs, cutting the vector fragments at such known restriction site, and isolating those fragments  
15 containing the selected marker.

16. A method of detecting messenger RNA transcripts produced by one or more genes  $G_t$  in a selected genomic structure in a test cell type, but not  
20 by the genes  $G_c$  contained in the corresponding genomic structure in a control cell type, where both the control and test cell genomic structures include at least about 10 genes which are active in producing messenger RNA transcripts, at different abundance levels, said method  
25 comprising

(a) providing a composition of control cell transcript species derived from the control cell genomic structure genes  $G_c$ , and present in the composition in substantially equal abundance,

30 (b) providing a composition of test cell transcript species derived from the test cell genomic structure genes  $G_t$ , and present in the test cell composition in substantially equal abundance.

-77-

hybridizing the control cell transcript composition with the test cell transcript composition, and

identifying those test cell transcript species  
5 which do not hybridize with control cell transcript species.

3 17. The method of claim 16, for use in detecting substantially all messenger RNA transcripts  
10 which are produced by the test cell, but not the control cell genomic structure, wherein the control cell and test cell transcript compositions each contain at least about one transcript species for each gene which is  
15 expressed by such control cell and test cell genomic structure, respectively, and in substantially equal molar abundance.

18. The method of claim 16, wherein the control cell transcript species are labeled with an  
20 affinity label which permits binding of the transcripts to a solid support, said hybridizing is carried out in the presence of excess molar concentration of control cell transcript species, and said identifying includes contacting the hybridized species with the solid  
25 support, and isolating hybridized species which do not bind to the support.

19. The method of claim 16, wherein the test cell transcript species are spotted individually on a  
30 filter, the control cell transcripts are radiolabeled, said hybridizing is carried out by hybridizing the radiolabeled species with the test cell species on the filter, and said identifying includes identifying those filter spots which do not contain radiolabel.

20. A method of detecting differences in the abundance of messenger RNA transcripts produced by one or more genes  $G_t$  in a selected genomic structure in a test cell type, with respect to the genes  $G_c$  contained in the corresponding genomic structure in a control cell type, where both the control and test cell genomic structures include at least about 10 genes which are active in producing messenger RNA transcripts, at different abundance levels, said method comprising

(a) providing a composition of control cell transcript species derived from the control cell genomic structure genes  $G_c$ , and present in the control cell composition in substantially equal abundance,

(b) transferring the composition onto two replica filters in which individual filter spots represent individual species of the composition,

(c) obtaining radiolabeled transcript species from the control and test cells, respectively,

(d) hybridizing the labeled control cell transcript species with one of the replica filters, and the labeled test cell transcript species, with the other filter,

(f) producing autoradiographs of each of the filters, and

(g) comparing the density of radiolabel associated with each spot on the test cell filter with that on the control cell filter, to determine the relative amounts of messenger RNA transcript associated with a gene in the genomic structure.

21. A method of preparing a clonal library composition of transcripts species which are (a) each derived from one of a plurality of genes  $G_i$ , in a cellular genomic structure, all of which genes are



active in a defined cell type in producing cellular messenger RNA transcripts, at various levels of messenger RNA abundance, and (b) present in substantially equal molar abundance, said method comprising

providing cellular transcript species derived from the different-abundance messenger RNA transcripts, preparing polynucleotide species which are homologous to the cellular transcript species, and present in substantially the same molar abundance as the homologous transcript species,

mixing the polynucleotide species with the cellular transcript species under conditions which promote hybridization between homologous polynucleotide and cellular transcript species,

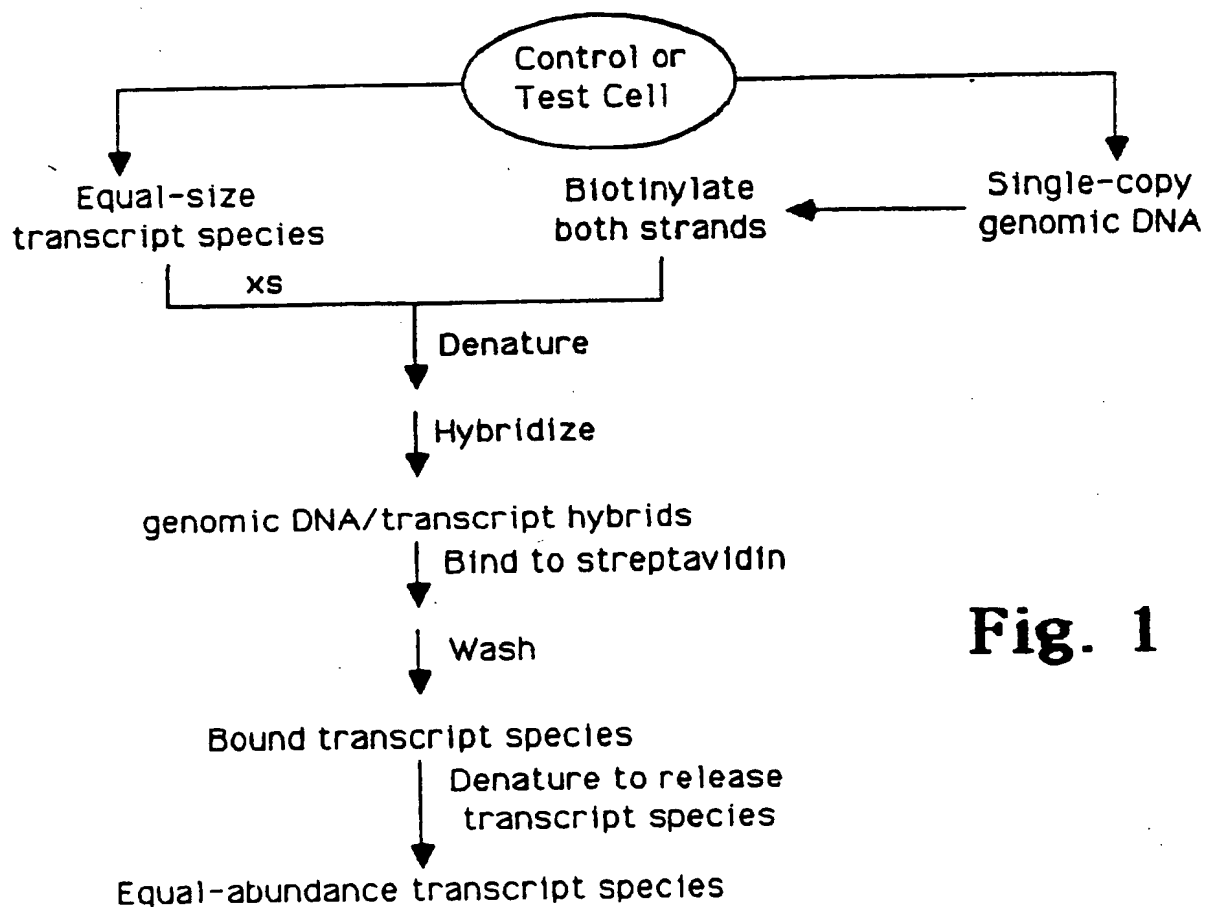
carrying out the hybridization until the transcript species which remain unhybridized all have substantially the same molar abundance,

separating hybridized from non-hybridized species, and

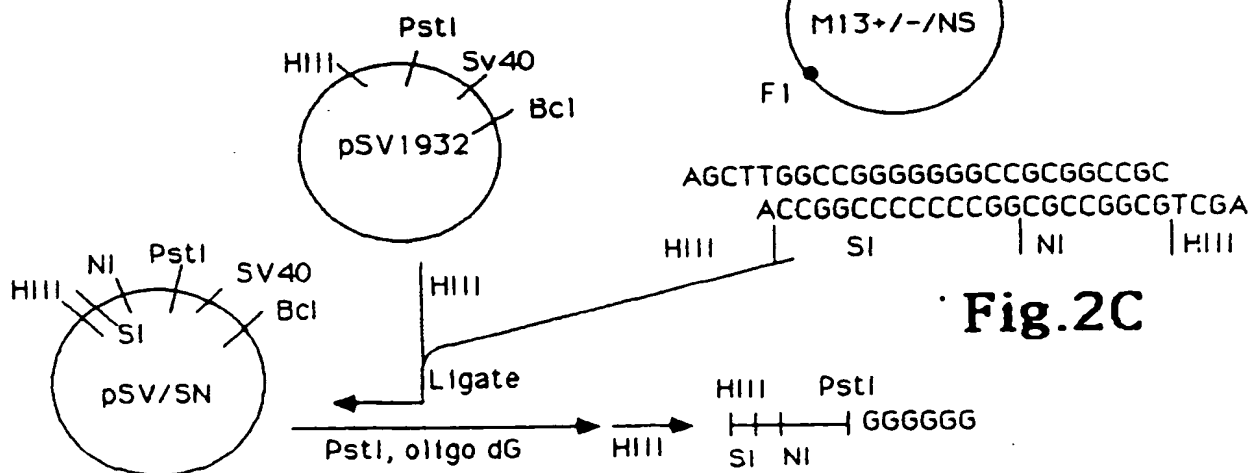
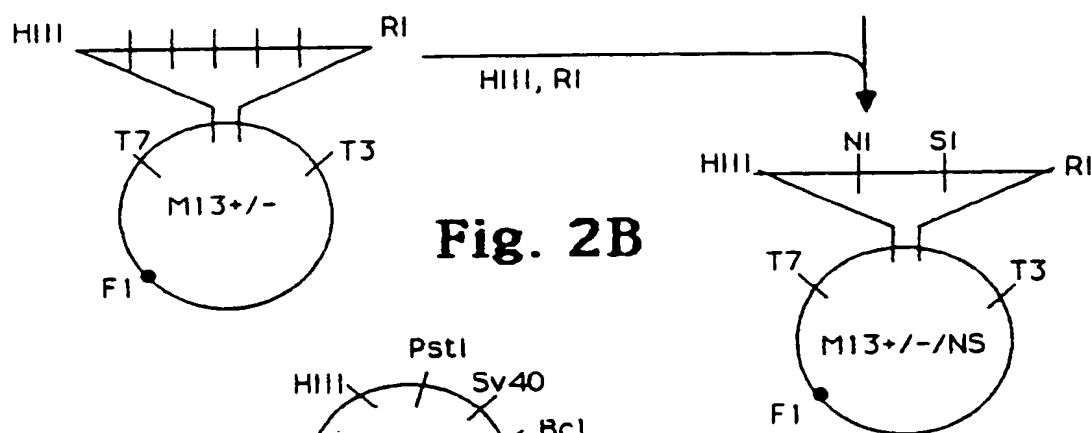
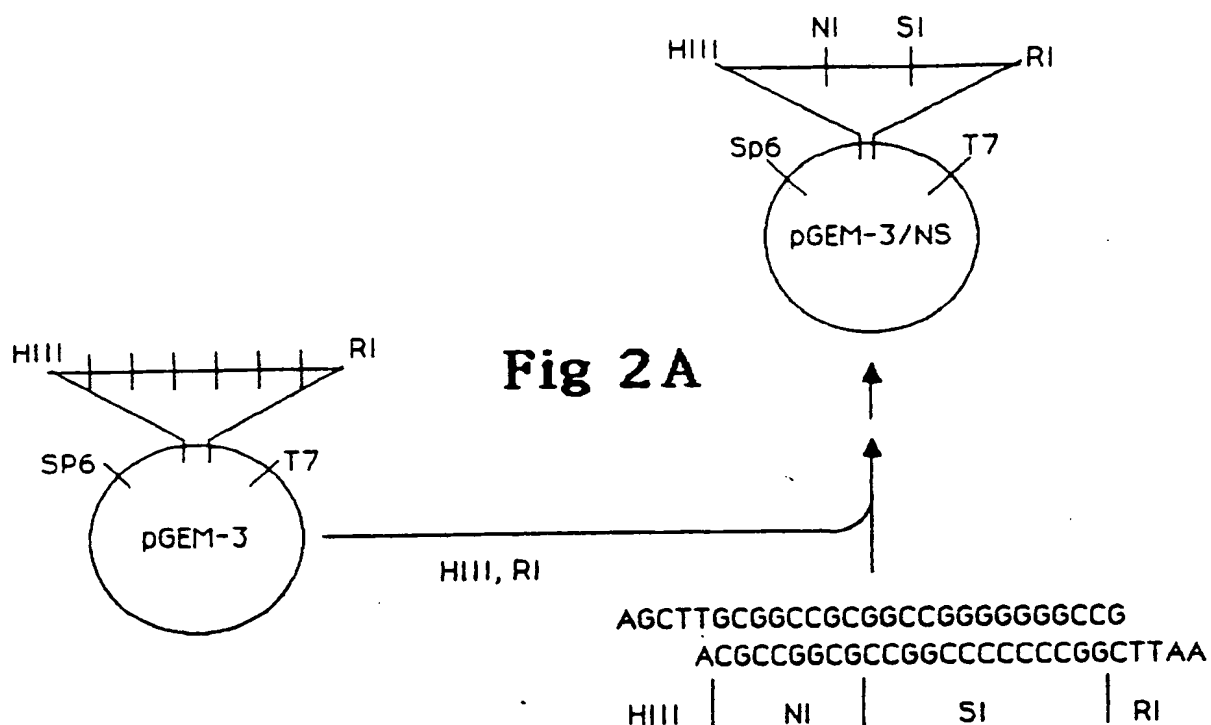
cloning the separated, non-hybridized species.

22. The method of claim 21, wherein the cellular transcript species which are provided and the polynucleotide species which are prepared are all within a defined size range less than about 1,000 nucleotides.

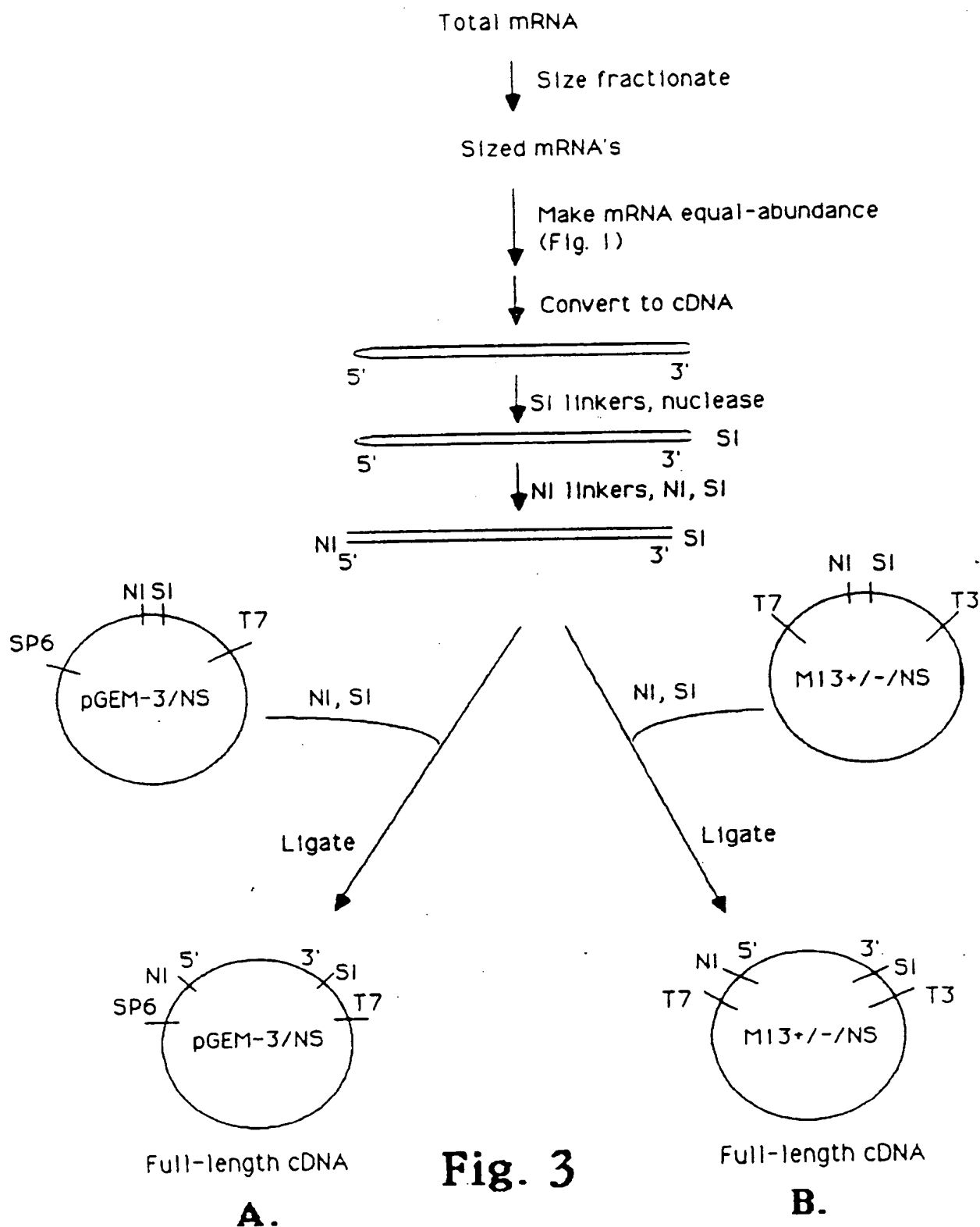
1 / 12

**Fig. 1****SUBSTITUTE SHEET**

2 / 12



3 / 12



4 / 12

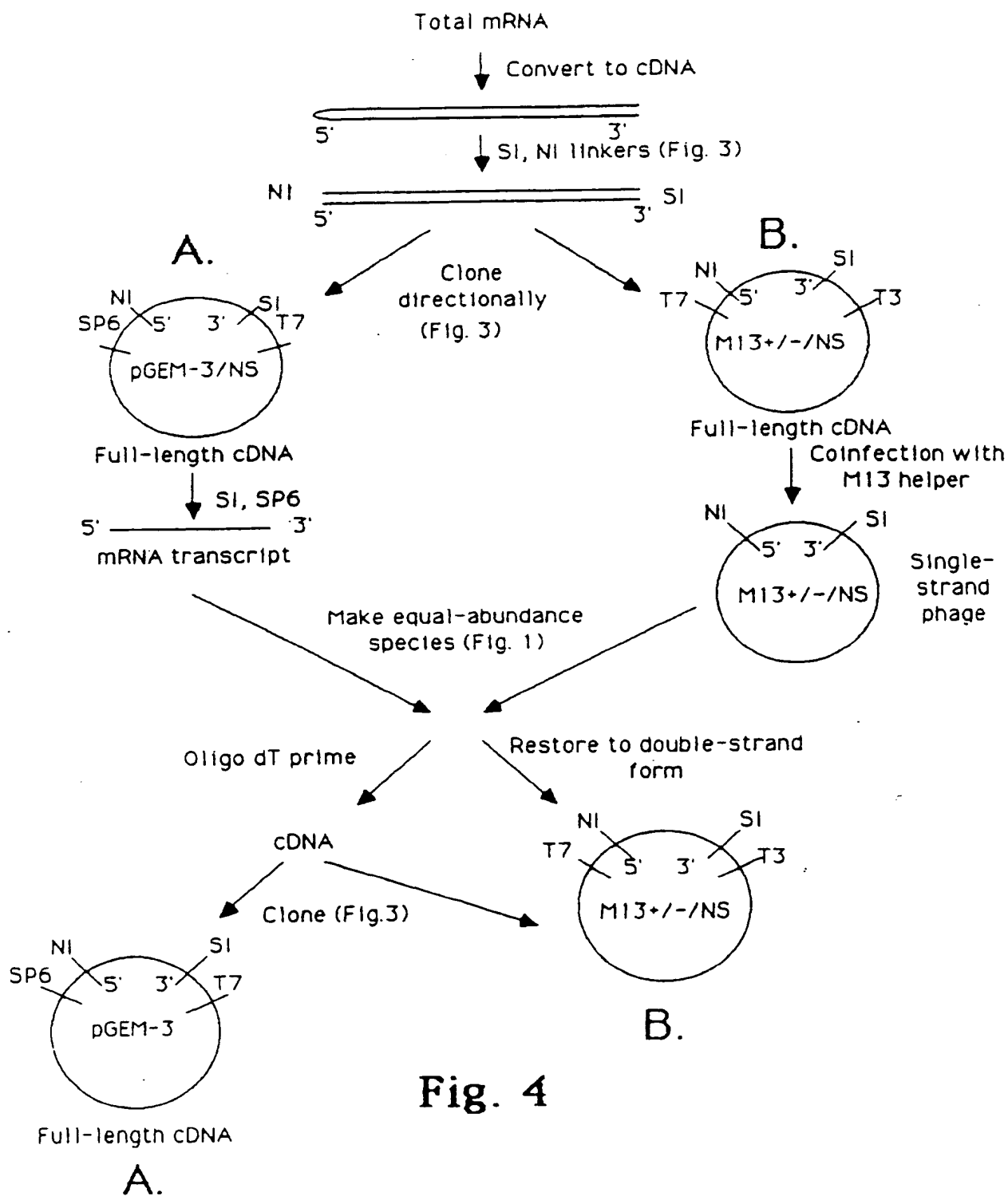
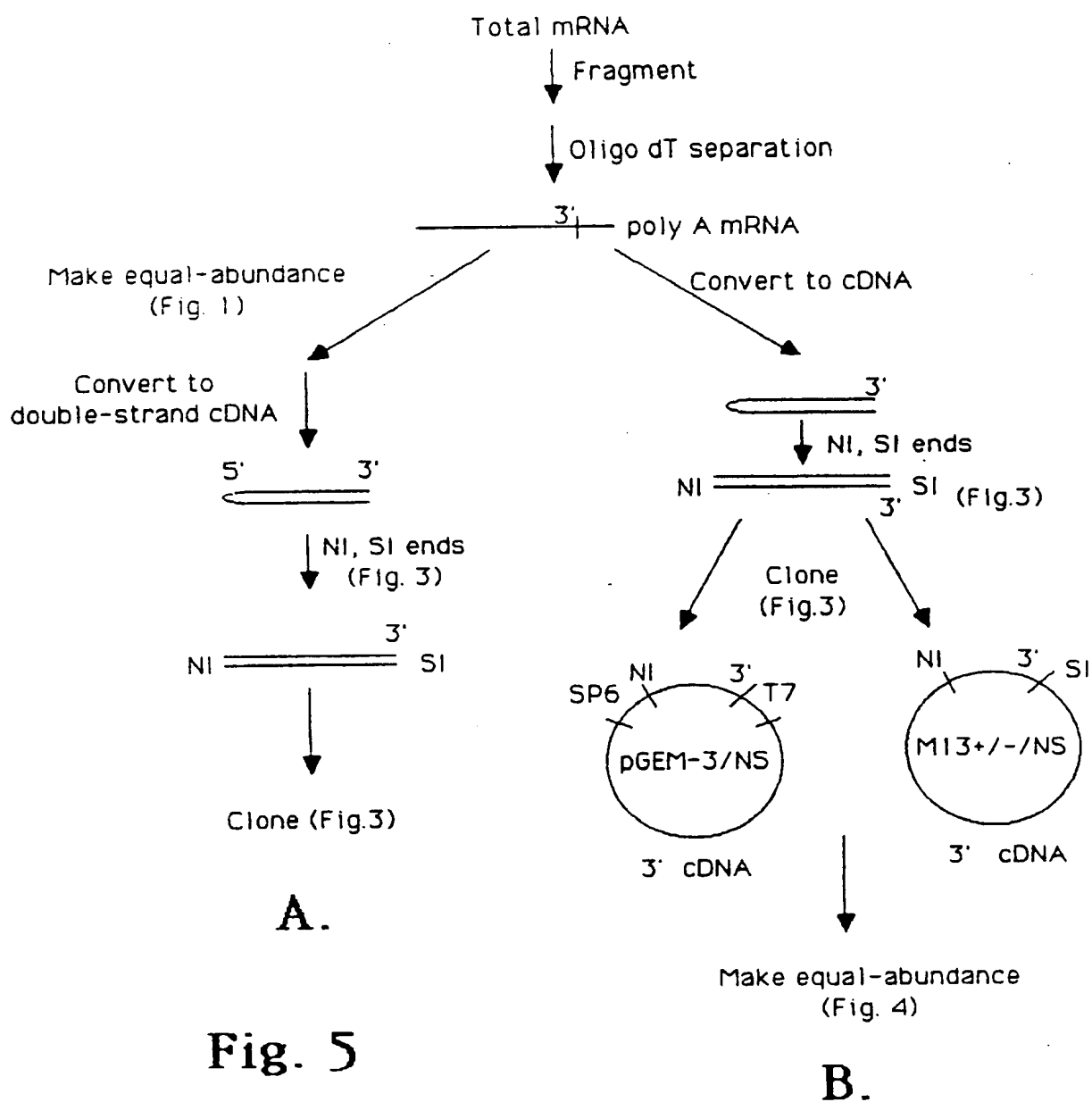


Fig. 4

5 / 12



6 / 12

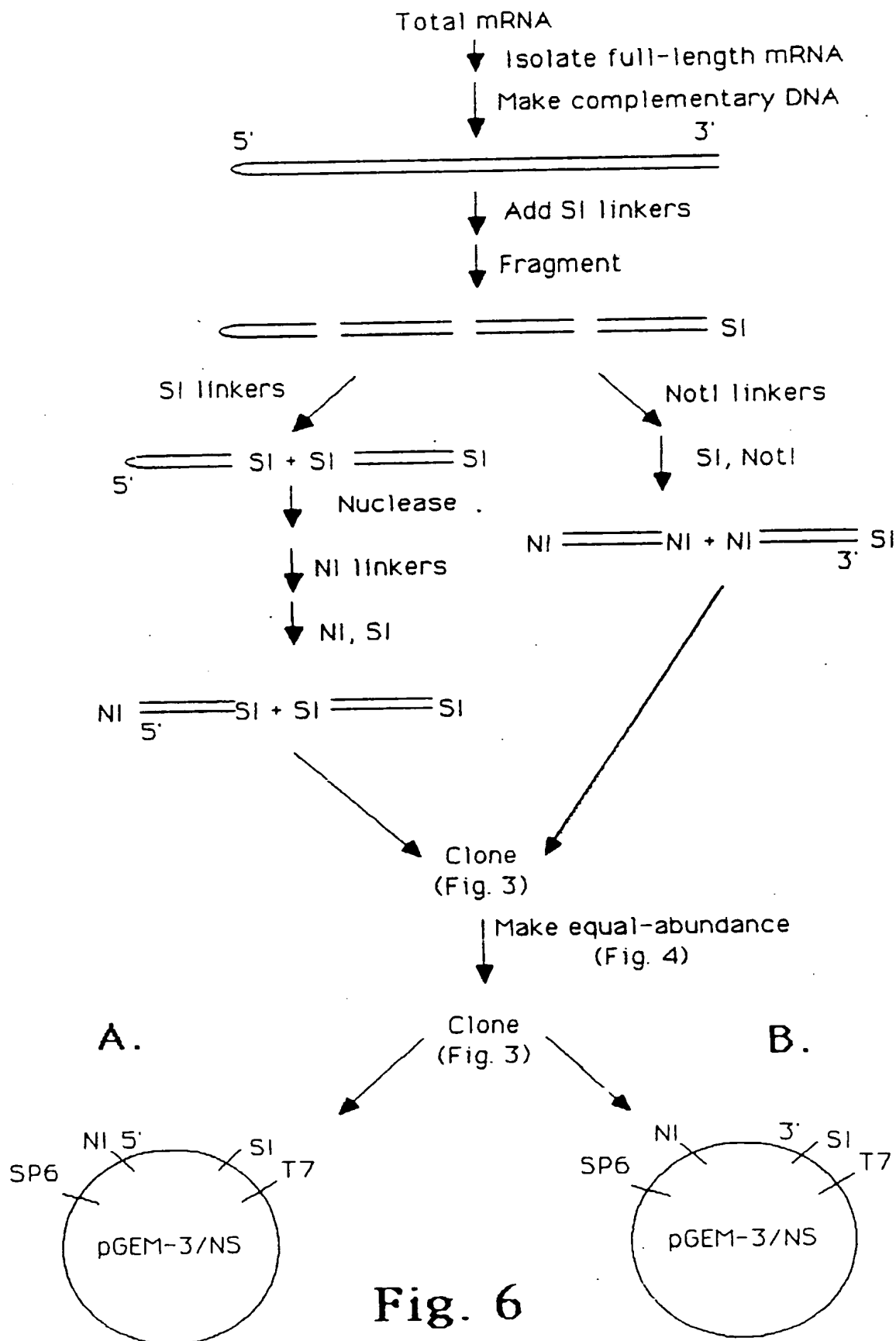


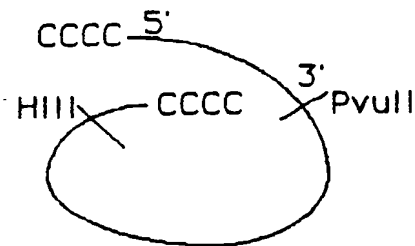
Fig. 6

7 / 12

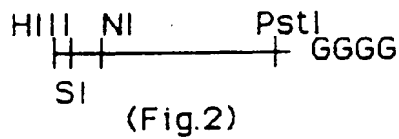
Full-length mRNA (Fig.6)

↓ Plasmid-primed synthesis

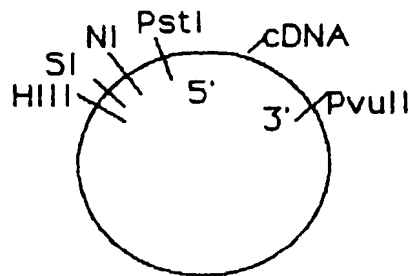
↓ Oligo dC tailing



↓ HindIII



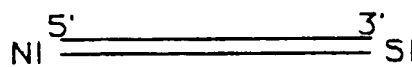
↓ Ligate



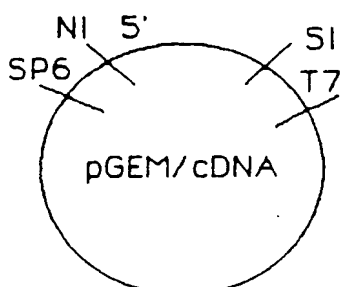
↓ Fragment

↓ Blunt-end, SI linkers

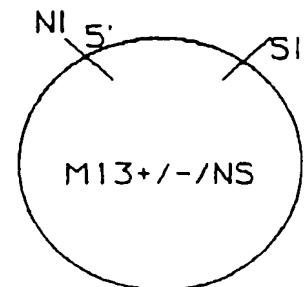
↓ SI, NI



Clone  
(Fig. 3)



5' cDNA

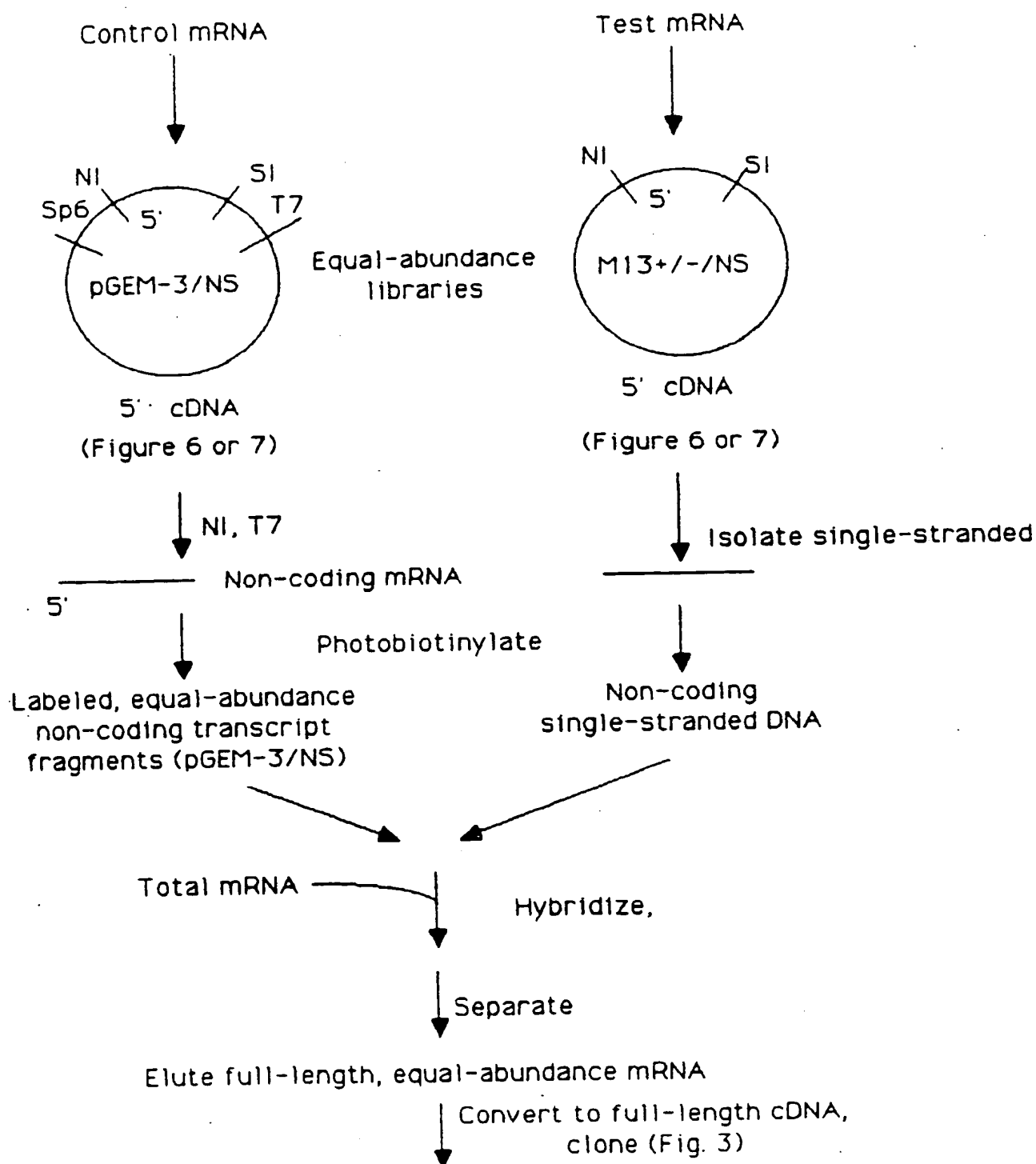


5' cDNA

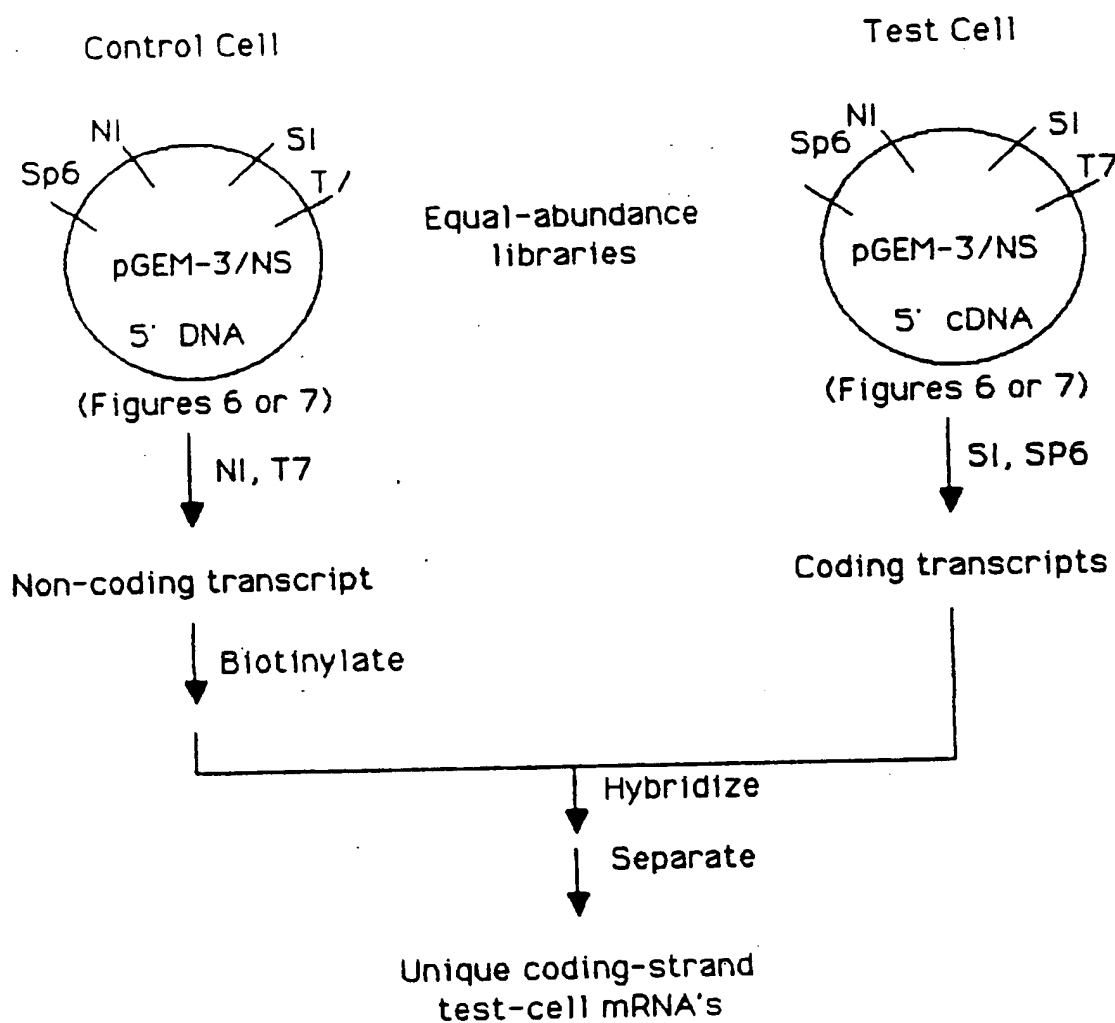
Make equal-abundance



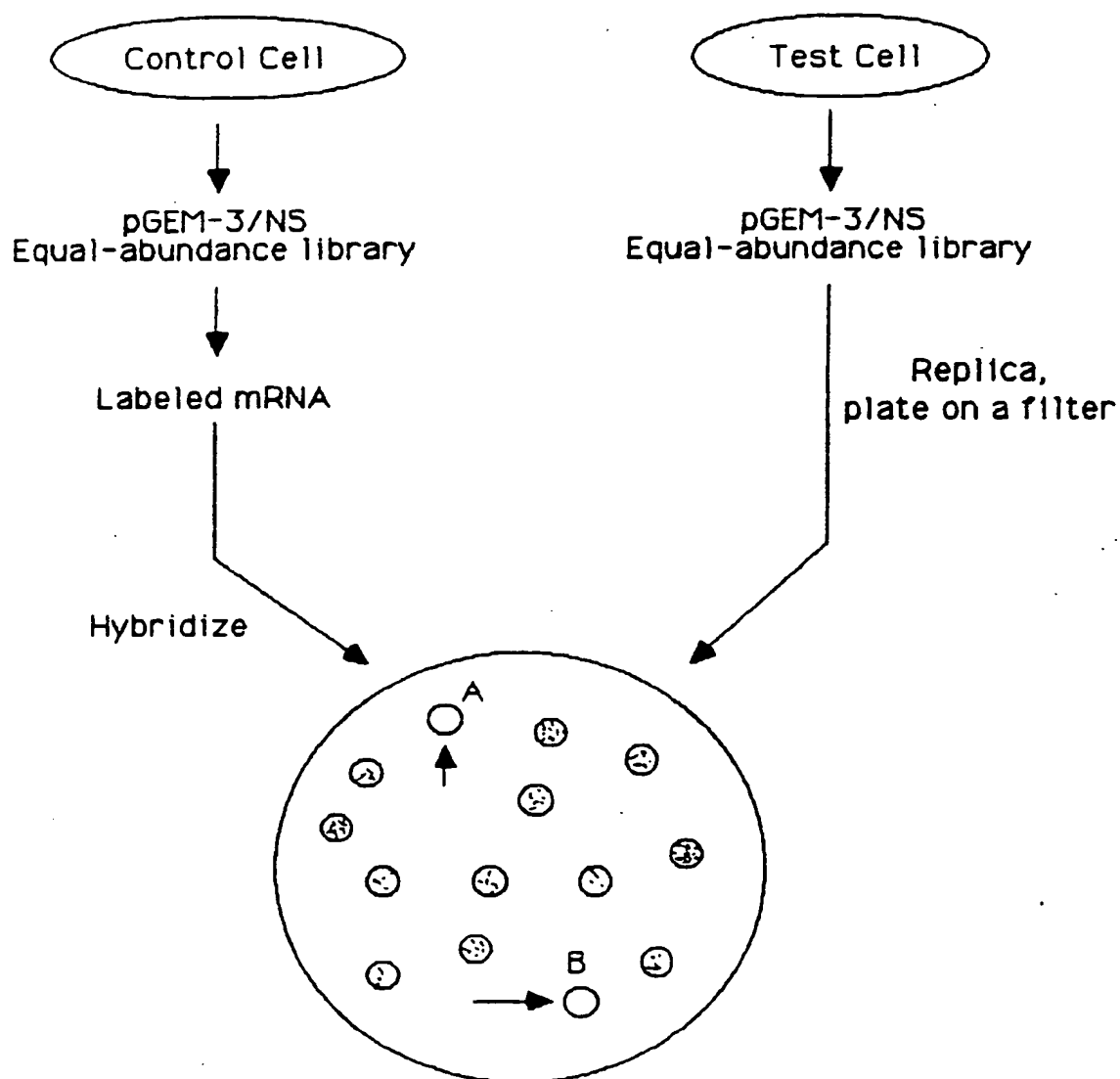
8 / 12

**Fig. 8**

9 / 12

**Fig. 9**

10 / 12



**Fig. 10**

11 / 12

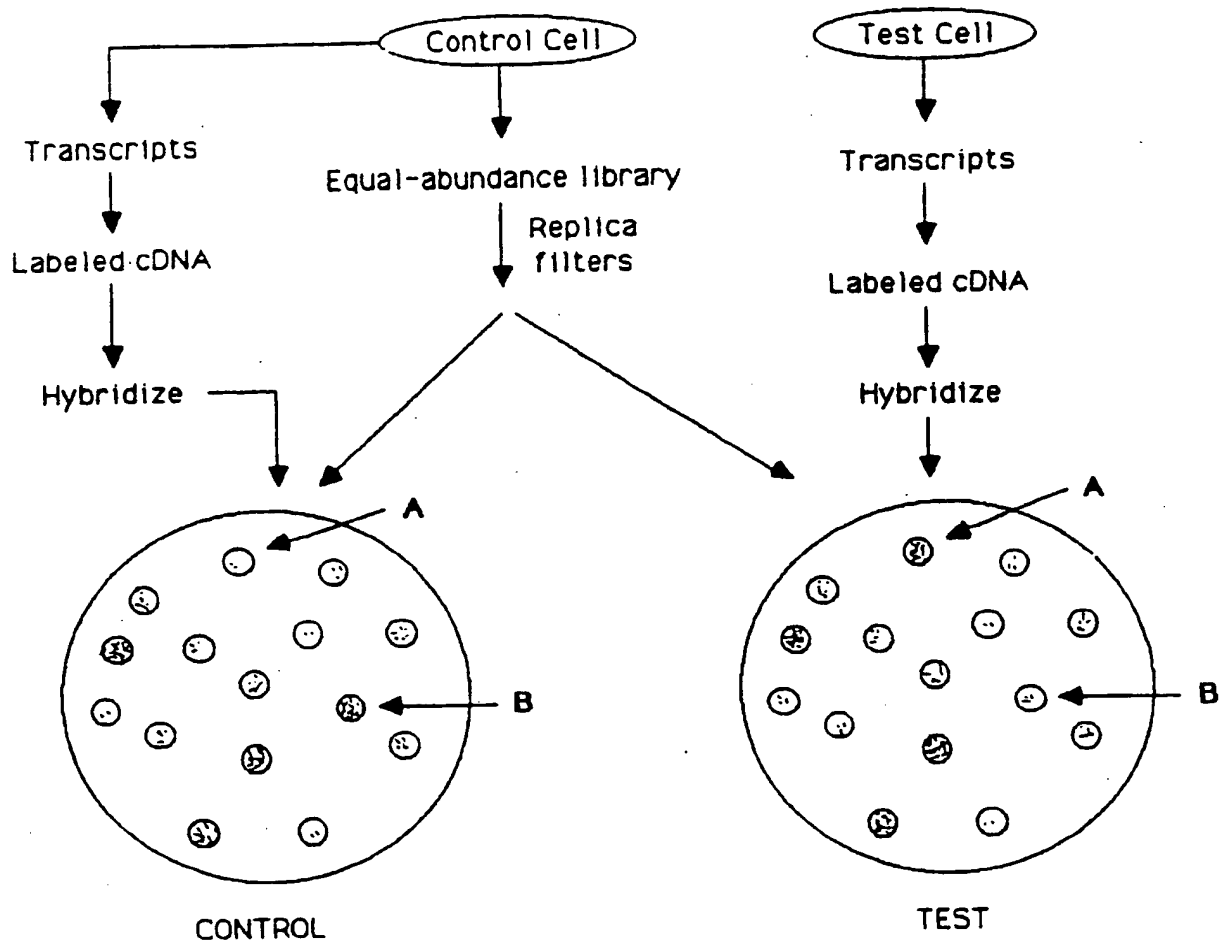
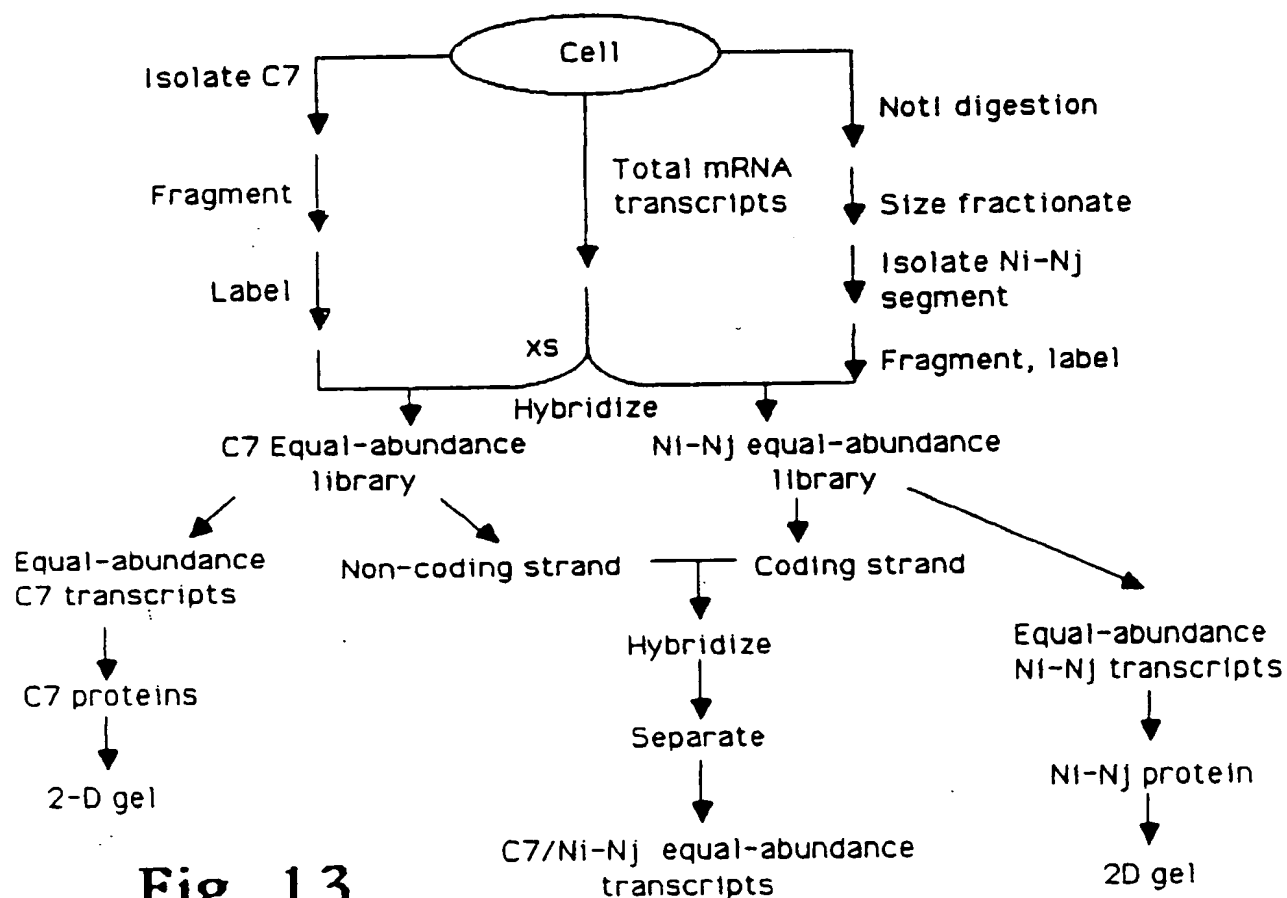
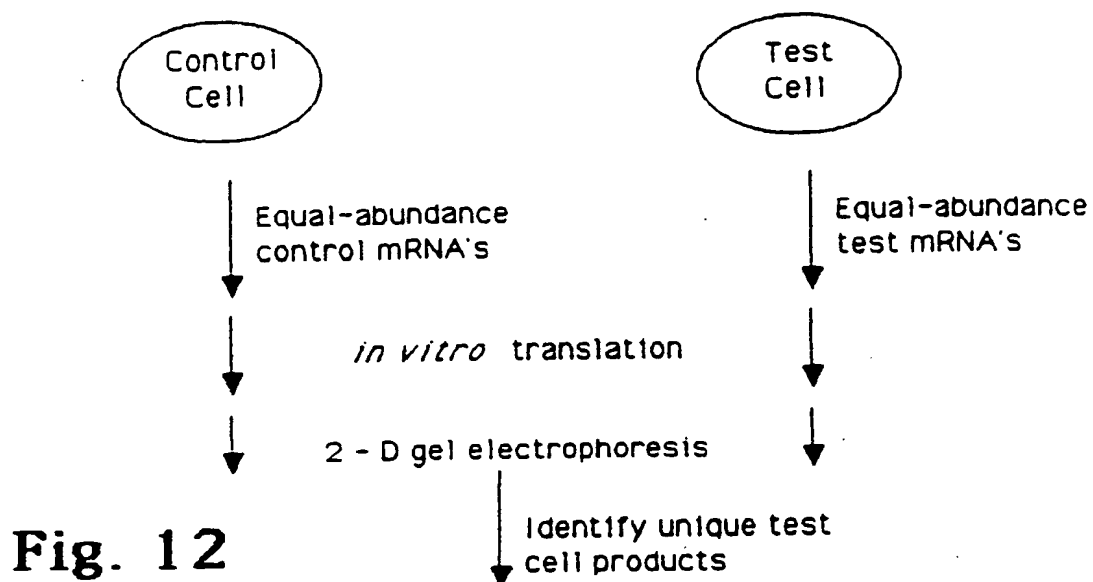


FIG. 11

12 / 12



# INTERNATIONAL SEARCH REPORT

International Application No. PCT/US88/01050

<b>I. CLASSIFICATION OF SUBJECT MATTER</b> (if several classification symbols apply, indicate all)		
According to International Patent Classification (IPC) or to both National Classification and IPC IPC (4) : C12Q 1/68; C12N 15/00; C07H 15/12 US CL : 435/6,172.3; 536/27; 935/78		
<b>II. FIELDS SEARCHED</b>		
Minimum Documentation Searched		
Classification System	Classification Symbols	
U.S.	435/6,172.3; 536/27; 935/78	
Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched		
COMPUTER SEARCH, CAS, BIOSIS, APS: MESSENGER, RNA ABUNDANCE		
<b>III. DOCUMENTS CONSIDERED TO BE RELEVANT</b>		
Category	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
Y	Miami Winter Symposium, Vol 16, issued 1979 (Philadelphia PA. USA), (PERRY ET AL), "The Processing of Messenger RNA And The Determination of Its Abundance", pages 187-206.	1-22
Y	Proc. Natl. Acad. Science, Vol. 79, issued July 1982 (Washington, D.C.), (KRAUS ET AL), "Purification of low-abundance messenger RNAs from rat liver by polysome immunoadsorption", pages 4015-4019.	1-22
Y	Gene, Vol. 35, issued 1985, (Amsterdam, Holland), (KOWALSKI ET AL), "Vectors for the direct selection of cDNA clones corresponding to mammalian cell mRNA of low abundance", pages 45-54.	1-22
* Special categories of cited documents:		
"A" document defining the general state of the art which is not considered to be of particular relevance		
"E" earlier document but published on or after the international filing date		
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)		
"O" document referring to an oral disclosure, use, exhibition or other means		
"P" document published prior to the international filing date but later than the priority date claimed		
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention		
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step		
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.		
"Z" document member of the same patent family		
<b>IV. CERTIFICATION</b>		
Date of the Actual Completion of the International Search		Date of Mailing of this International Search Report
01 July 1988		29 JUL 1988
International Searching Authority		Signature of Authorized Officer
ISA/US		Alvin E. Tanenholtz

## III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)

Category *	Citation of Document, <sup>14</sup> with indication, where appropriate, of the relevant passages <sup>17</sup>	Relevant to Claim No <sup>18</sup>
Y	BENJAMIN LEWIN, <u>GENES</u> , second edition, Published 1985 by John Wiley & Sons (New York USA) pages 304-311.	1-22
Y	US, A, 4,394,443 (WEISSMAN ET AL), Published 19 July 1983. See entire document.	1-22
Y,P	US, A, 4,675,285 (CLARK ET AL), Published 23 June 1987. See entire document.	1-22
Y	<u>Chemical Abstracts</u> , Vol. 102, No. 21, issued 27 May 1985, (Columbus, Ohio, USA), (HORWICH ET AL), "Strategies for the molecular cloning of low abundance messenger RNAs", see page 154, Col. 2, the abstract No. 180176t, Mol. Basis Lysosomal Storage Disord. 1984, 365-85 (Eng).	1-22
Y	<u>Chemical Abstracts</u> , Vol. 105, No. 17, issued 27 October 1986 (Columbus, Ohio, USA), (RHYNER ET AL), "An efficient approach for the selective isolation of specific transcripts from complex brain mRNA populations", page 358 Col. 2, the abstract no. 149268n, J. Neurosci. Res 1986 16(1) 167-81 (Eng).	1-22
Y	<u>Chemical Abstracts</u> , Vol. 105, No. 25, issued 22 December 1986 (Columbus, Ohio, USA), (SHELLY ET AL), "The use of complementary RNA and S1 nuclease for the detection and quantitation of low abundance mRNA transcripts" page 396 Col. 2, the abstract no. 222148e, Bio Techniques 1986, 4(5), 434-8 (Eng).	1-22
Y,P	<u>Chemical Abstracts</u> , Vol. 107, No. 17, issued 26 October 1987 (Columbus, Ohio USA), (KIM ET AL), "Characterization of cytoplasmic RNA containing mRNA encoding the 70,000-dalton heat stress protein", page 167, the abstract No. 148339r, Korean J. Genet. 1986, 8(4) 221-9 (Eng).	1-22

**THIS PAGE BLANK (USPTO)**